



Article

EmoSpell, a Morphological and Emotional Word Analyzer [†]

Maria Inês Maia  and José Paulo Leal * 

CRACS & INESC-Porto LA, Faculty of Sciences, University of Porto, 4099-002 Porto Porto, Portugal; up201101593@fc.up.pt

* Correspondence: zp@dcc.fc.up.pt

[†] This paper is an extended version of our paper published in 6th Symposium on Languages, Applications and Technologies (SLATE 2017).

Received: 29 September 2017; Accepted: 7 December 2017; Published: 3 January 2018

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 to also 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. Additionally, this paper details the Web application and Web service that exploit this dictionary. It also presents 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.

Keywords: sentiment analysis; opinion mining; Emotion API

1. Introduction

Sentiment analysis, also known as opinion mining, can be classified as the identification of opinions, emotions, and evaluation in texts [1] toward topics, events, entities, or individuals. The use of dictionaries containing words annotated with semantic or polarity orientation is a frequent approach to perform this kind of analysis. These dictionaries can be created manually or by using lemmas, thus automatically expanding the list of words [2].

This paper describes the implementation of EmoSpell (<http://handspy.up.pt/EmoSpell/>), an emotional word analyzer for the Portuguese language. Some content of this paper can be found in the published conference paper “An Emotional Word Analyzer for Portuguese” [3]. Comparing with the already published paper, here we provide more details about the motivation and the lexical bases used to create EmoSpell. EmoSpell contains a dictionary with semantic annotations related to the emotional value of words. It is based on a morphological analyzer named Jspell [4] and on EMOTAIX.PT [5], 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.

The motivation for EmoSpell comes from the research on the cognitive processes in writing, more specifically HandSpy (<http://handspy.up.pt/>). This platform is a Web-based application that focuses on these processes, as a collaborative environment for researchers to study and analyze text

productions made by children. One of the most important tasks carried out by HandSpy is to find the relationship between pauses in the writing and their cognitive processes.

The integration of EmoSpell would be a great improvement for the HandSpy analysis of linguistic and emotional texts. The fact that EmoSpell provides the syntactic and emotional categories of words and information about the dominant category of a text, as well as the number of emotional words, would allow a new object of study in this research on cognitive processes in writing.

The creation of the emotional word analyzer EmoSpell comes with two interfaces—a Web graphical user interface and an application programming interface (API). The latter was created with the aim of being integrated in other applications. Thus, HandSpy can integrate this analyzer and allow the annotations with the morphological and semantic analysis of words, as well as the emotional value that they represent.

Jspell uses a dictionary of lemmas with a set of morphological rules, avoiding 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 in order 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. For instance, if the lemma “sentir” (“to feel”, in Portuguese) is annotated with “benevolência” (“benevolence”), “afeição” (“affection”), and “amor” (“love”) as emotional categories, so are other forms of this verb, such as “senti-me” (“I felt”). It is also possible to reverse the emotional category of a word when a negative prefix is added, such as “in”, “des”, “de”. For example, for the words “feliz” (“happy”) and “infeliz” (“unhappy”), it will be annotated that this emotional classification is the opposite.

EMOTAIX.PT is a lexical database with 3992 collected words. These are classified according to their valence (that is, positive, negative, and neutral) and semantic nature, which consists of a hierarchical classification of emotional categories.

There are other Portuguese lexicons with sentimental annotations, such as SentiLex-PT [6]. This one in particular provides only word polarity, through a positive, negative, or neutral classification. Therefore, EmoSpell will allow a more detailed analysis of sentiments for the Portuguese language.

1.1. Cognitive Processes

In order to collect information for the analysis of these processes, researchers focus on several indicators, such as the eye movement during writing [7] and the different stimuli, observing aspects such as velocity, movements, and duration of the writing production [8]. These two last aspects are part of the analysis carried out by HandSpy.

1.1.1. HandSpy

The project “Desenvolver, Automatizar e Autorregular os Processos Cognitivos na Composição Escrita” (DAAR) is a research project being developed at the Faculty of Psychology and Education Sciences of the University of Porto, was created for the research on the cognitive processes in writing. HandSpy was developed to support this research. The main goal of this research is to determine the factors that influenced the development of writing skills, based on cognitive processes in writing. The study objects are written productions composed by school children.

HandSpy is a collaborative environment platform that aims at the management of experiments in the study of cognitive processes in writing [9], which covers all the mentioned experiment processes. Therefore, this platform is a Web-based application that relies on the digital collection of writing and data analysis. By using a Web server for data storage, various researchers are able to share a common repository in order to work simultaneously in an experiment, with the proper management and analysis of the collected data.

This data collection is performed through a Livescribe Smartpen (<https://www.livescribe.com/pt/smartpen/>), a digital pen containing an LCD display with a navigation menu for information status and an infrared camera to record the handwriting. Then, a system based on *Java Micro Edition* runs

and works together with a micro dotted paper, which consists in printed microdots on its surface to capture the position in the paper, containing the strokes and timestamps of each point.

The data is obtained by these digital pens and stored in the database. The Ink Markup Language [10] format is a W3C recommendation created with the aim of storing and representing the collected data in digital ink. InkML is an XML [11] data format consisting of a set of strokes that define handwriting characteristics such as pen orientation while writing, width, color, and timestamps, among other important information used by researchers in the data processing. An InkML file contains the properties of a written text, more specifically the set of strokes that represent a sequence of continuous ink points, with the X and Y coordinates of each point. It can have the correspondent timestamps as well.

HandSpy is a Web-based application that follows a three-tier architectural model consisting of presentation, logic, and data tiers. These lie on the Web interface, on the Web server, and in the XML database, respectively.

When it comes to the application design, HandSpy's layout contains three main tabs, plus one for the administrative tasks. Figure 1 demonstrates the platform's design and its respective tabs. Each one of them has a specific function:

Project Tab is where basic information of the project is shown, containing buttons to manipulate events. This allows the creation of tasks and the addition of participants.

Upload Tab has as main function the upload of InkML files extracted from smartpens. It then uploads the protocols, assigning them one task and one participant.

Analysis Tab is where protocols are selected and analyzed. So, as shown in Figure 1, where the Analysis Tab is selected, the data of the selected protocol is in the workspace on the left side (bursts, pauses, and distance). On the right side is the image of the text to be analyzed.

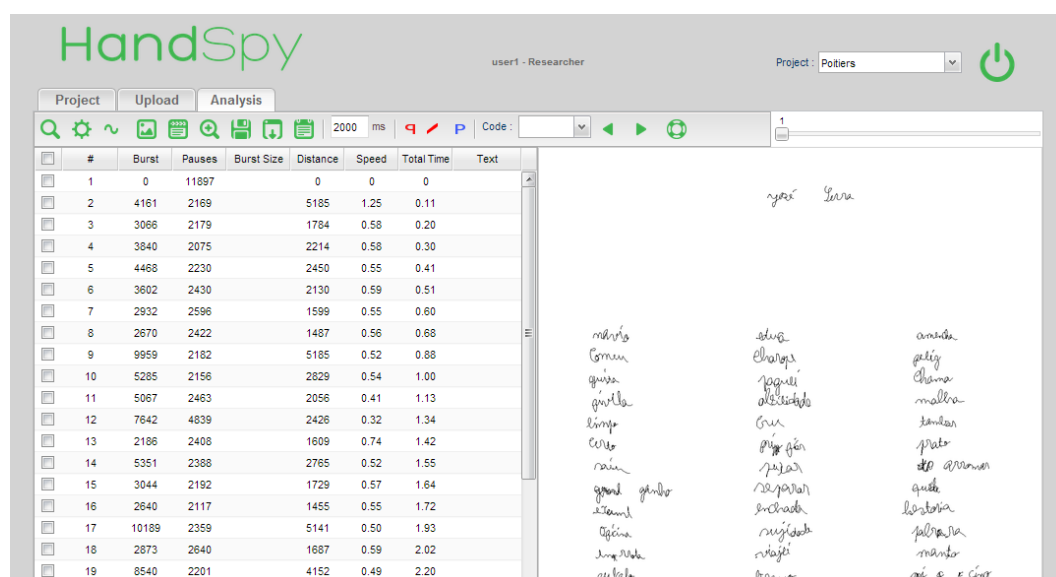


Figure 1. Handspy—Analysis Tab.

In the image of the selected text of Figure 1 there are three annotations, being “p” the beginning of the burst and “q” the end of the latter .

These were created with the purpose of graphically representing the data and to provide the investigator with a better perception of the information. Both contribute greatly to the study of cognitive processes in writing.

1.2. Sentiment Analysis

One of the sentiment analysis approaches is the creation of lexicons to identify sentiments in texts and words. That can be done automatically, based on a set of lemmas and through methods to create flexed forms, as it can be observed on Jspell and SentiLex-PT. This subsection explains the importance of the analysis of sentiments and how this can be accomplished. It also presents some lexicons and sentiment analysis API.

1.2.1. Applications and Challenges

There has been a considerable increase in studies regarding sentiment analysis in texts, becoming very common from 2000 onwards [12]. This analysis might play an important role in other technologies (e.g., recommendation systems), and can also provide information for social studies and marketing.

The availability of information from microblogging Web-sites users turns it into the most famous source of data in this area. Application-oriented research papers on microblogging Web-site such as Twitter are based on consumers' opinions of products and services, which is valuable information for companies [13]. It is also useful to gauge the political opinion of the population on a certain subject [14].

Opinion analysis has always been a relevant topic, either for individuals or for businesses and organizations [12]. An example of this is when a company wants to know if its customers are satisfied with some product or if the reviews from a release or campaign are positive or negative. This analysis can be used for statistical purposes as well. For example, if there is the need to know bloggers' and twitterers' attitudes regarding the election of the president of the United States.

As mentioned before, the use of sentiment analysis can be a valuable application for recommender systems. This kind of system is based on information filtering, in order to predict the user preferences and classifications for an item. With the use of a sentiment analysis such as an opinion of a movie in IMDb (<https://www.imdb.com>), it is possible to extract the sentiment and the strength of opinions in words provided by the user. Consequently, with the use of techniques like collaborative filtering [15], it is possible to infer personalized recommendations according to some of their preferences.

These can provide information beyond the usual "good" or "bad" reviews. If the company wants to know why the product has not been purchased as expected, sentiment analysis can answer this question by providing the motives such as "very expensive" and "poor design".

Through words, an individual can express their liking or disliking towards a certain matter, thus expressing positive or negative feelings. On the one hand, if a sentence contains the words "like", "approved", or "good", it can be concluded that the individual's positions towards the matter is possibly positive. On the other hand, if the words are "hated", "disapproved", or "bad", it is expected to be a negative sentiment.

However, in the process of sentiment analysis, there are some challenges when it comes to perceiving the accurate meaning of the author. Words can be used with different meanings and can convey different emotions. It is not as simple as classifying some words as positive and others as negative.

When writing a text or a review, there is the possibility of the writer using words of an opposite emotional category from the context of the text. When classifying single 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 the positive word "great" [16].

It is important to focus on the fact that people do not express their opinions as the annotated lexicon wishes they did. Individuals can be contradictory in their sentences in order to enhance their feelings. For example, the use of sarcasm in sentences is a barrier to the emotional analysis of texts. In the sentence "I absolutely adore it when my bus is late", the positive word "adore" is used sarcastically to highlight a negative situation [17].

Another challenge is when the user is writing about some event or history and uses words that can be interpreted as some emotional classification. For example, the sentence "The war has created

millions of refugees”—the writer is merely describing a situation but, as he uses heavier words, from an emotional charge point of view, this can be classified as a negative word or sentence [18].

Thus, there are still some challenges to be taken into account when creating a sentiment lexicon. Incorporating techniques such as multiword analysis or identification of the polarity target makes a more effective analyzer.

1.2.2. Sentiment Lexicon

Sentiment analysis has two approaches: classifier-based and lexicon-based. The first uses machine learning techniques, a problem of text categorization, and the latter uses sentiment lexicons to analyze the text and perceive the sentiments that it contains.

Additionally, there are two main approaches when constructing a sentiment lexicon: manual and automatic. In manual creation, there is no complexity in the algorithmic point of view. It is based on the enumeration of a list of words and the annotation of the sentiment classification. This annotation is made by humans so, besides the probability of human errors, the creation time and lexicon size are also not as expected. The automatic creation of a sentiment lexicon is made by creating a set of lemmas and then using rules or methods to expand them.

Sentiment lexicons were created to analyze texts emotionally. They list words with an emotional classification and provide information regarding their emotional load when present in a text. This emotional load can be represented by polarities (categorizing words by positive, negative, or neutral), by values representing the emotion strength, or by emotional categories.

There are some lexicons already created for some languages. For the Portuguese language, the SentiLex-PT will be explained in more detail in this paper, since it is a part of the EmoSpell creation. To better understand the differences between the lexicons, some of these are briefly explained below:

- General Inquirer (GI)—created in 1966, it is a content analysis tool that consists of a manually created database of words. It is used to count words of emotional categories and it has 182 categories, each one being a list of words and word senses, combining “Harvard IV-4” and “Lasswell” dictionaries. It is possible for the user to add more categories [19].
- Linguistic Inquiry and Word Count (LIWC)—a text analysis software program that calculates the degree of use for different categories of words across a wide array of texts. It uses the lexicon resource (the LIWC dictionary of words and word stems), each being filed into one or more sub-dictionaries. It classifies words in psychologically-relevant categories [20].
- EMOTAIX—a similar tool to LIWC, but for the French language. [21] An EMOTAIX adaptation was also created for the Portuguese language—EMOTAIX.PT [5]. EMOTAIX has the same levels and categories as EMOTAIX.PT.
- SentiWordNet—an extension of WordNet. WordNet uses an English dictionary containing nouns, adjectives, verbs, and adverbs that can be called “synsets”. “Synsets” are sets of cognitive synonyms, linked by semantic and lexical relationships. With this linking of words, Wordnet groups them based on their meanings. [22]. SentiWordNet added three sentiment values to each “synset”. With this addition, the lexicon assigned three scores, and thus, each “synset” has positivity, negativity, and objectivity [23].
- Sentiment Orientation CALculator (SO-CAL)—a system that contains a dictionary of annotated words with semantic orientation. SO-CAL has 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 [2].

1.2.3. Sentiment Analysis API

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

- TweetSentiments (<https://www.mobomo.com/2010/11/sentiment-analysis-using-tweetsentimentscom-api/>)—returns the sentiment of Tweets and is based on 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. It uses the LIBSVM library for SVM, which is implemented in C++ and also offers a Ruby and Rails implementation. This REST API can be used as a Web Service or as a standalone application, and the response format is in JSON.

- Text Processing (<http://text-processing.com/>)—an API with JSON over HTTP Web service. It is free and open-source. It performs a phrase extraction, sentiment analysis, part-of-speech tagging, and named entity recognition.
- ML Analyzer (<http://mlanalyzer.sudo.me/>)—provides several text analyses, including feelings, text classification, language detection, locations extractor, adult content analyzer, and article summarization. It contains a REST API with JSON response format.
- WebKnox Text-Processing(<http://webknox.com/>)—natural language processing of texts such as the determination of the feeling, identification of the language, classification of the quality of writing, auto-correction of a text, extraction of data and locations, and tagging of a text with part-of-speech tags. The response format is in JSON.
- Skyttle (<http://www.skyttle.com/>)—provides services to extract patterns from the text such as sentiment terms, constituent terms (meaningful expressions), and entities such as names of people, places, and things. Supported languages are English, French, German, and Russian. Is a SaaS (Software as a service) system that offers an option to receive the text with XML-annotated keywords and sentiment.
- nlpTools (<http://nlptools.atrilla.net/web/>)—text classification and sentiment analysis for natural language. It is an API focused on online news media. It is written in the PHP programming language and the API is JSON over HTTP RESTful Web service.
- Yactraq Speech2Topics (<https://yactraq.com/>)—converts audiovisual content into topic metadata. This conversion is done through speech recognition and natural language processing. It has a REST API with JSON response format.

2. Results

The validation of the proposed approach compared the results obtained in the analysis 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 three texts describing: a traumatic moment, a happy experience, and their daily routine.

The validation with EMOTAIX.PT used a simple program, recognized by the SentiLex-PT analyzer, that counts the words of the text. In order to compare the EMOTAIX.PT with the EmoSpell, some of the texts were used to verify the values that EmoSpell provided.

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 to EMOTAIX.PT and EmoSpell results, respectively. It lists results of the analysis of three different tasks: one positive, one negative, and one neutral for each participant. To represent these texts, which can be verified in each line of the table, each text is classified by the participant number and by the emotional type of the text written. For example, the first line “1-Pos” represents the positive text of participant number one. For each text, it lists the number of emotional words given by EMOTAIX.PT and EmoSpell, distinguishing the positive, negative, and neutral words. As can be seen in all texts, EmoSpell can detect a larger number of emotional words.

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

	Emotional			Positive			Negative			Neutral		
	EX	ES	$\Delta\%$	EX	ES	$\Delta\%$	EX	ES	$\Delta\%$	EX	ES	$\Delta\%$
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

If we observe the positive text from the third participant (the line 3-Pos), it is possible to verify that EMOTAIX.PT detected 10 emotional words and EmoSpell 17 words. Furthermore, it detected that it is a positive text considering the large number of positive words and the absence of negative words.

As mentioned before, EmoSpell was also validated against SentiLex-PT. The developed program uses the SentiLex-PT file to count the number of emotional words and their type, and to compare the values detected in this analyzer with EmoSpell. 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 to SentiLex-PT and EmoSpell, respectively.

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

Participant—Texts	Words											
	Emotional			Positive			Negative			Neutral		
	S	ES	$\Delta\%$	S	ES	$\Delta\%$	S	ES	$\Delta\%$	S	ES	$\Delta\%$
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

3. Discussion

One of the aims of the experiment mentioned in the Results section was to verify the possibility of categorizing between positive, neutral and negative texts. From the results of 81 written texts obtained by university students, it was verified that they write more emotional words in the negative and positive condition than in the neutral condition. Additionally, 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 this ability to synthesize the polarity of texts. Generally, all participants used at least one emotional word of the opposite category.

The existence of a large number of words of contrasting polarity—particularly in negative texts—has several motives. The high number of positive words in the negative and neutral texts can be explained by the way EMOTAIX.PT categorizes words. Words such as “boyfriend”/“girlfriend” and “friend” are defined in EMOTAIX.PT as positive. When participants write about a specific experience, they typically use these words to refer to the persons involved, without an emotional connotation.

Another explanation is 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 it happens with EMOTAIX.PT, we can verify in the Table 2, that EmoSpell detects a greater number of words than SentiLex-PT. In this case, the divergence of the average number of words detected is actually higher.

Besides the positive difference between the SentiLex-PT and EMOTAIX.PT with the EmoSpell, the latter also gives information about the dominant category of each text and the classifications of each word. This is important when the classification of a text by a top category is desired.

The emotional word detection was based not only 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.

Contrasting with the EmoSpell and EMOTAIX.PT comparison, SentiLex-PT can verify a large number of emotional words, and in some cases SentiLex-PT can even detect 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. These words are currently being categorized by the authors of EMOTAIX.PT, and they will be available on the next version of EmoSpell.

Nevertheless, 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 detects 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 were *“benevolence-affection-love”*, and *“non-specific emotions”*. The three positive texts have the same dominant categories as the negative ones. This can be explained by the fact that the participants used words such as “friend”, “girlfriend/boyfriend”, “to feel/feeling”, and words of the *“benevolence”* category several times. 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.

To conclude, it is possible to verify that EmoSpell can, on average, detect more words than the two emotional dictionaries on which it is based. To verify this, Figure 2 represents a graph that summarizes this comparison. The average number of words detected in these texts is sometimes double or triple. The fact that EmoSpell expands a dictionary with rules to generate more words with the emotional categories from EMOTAIX.PT is the main reason for these figures and for the better detection of Portuguese emotional words.

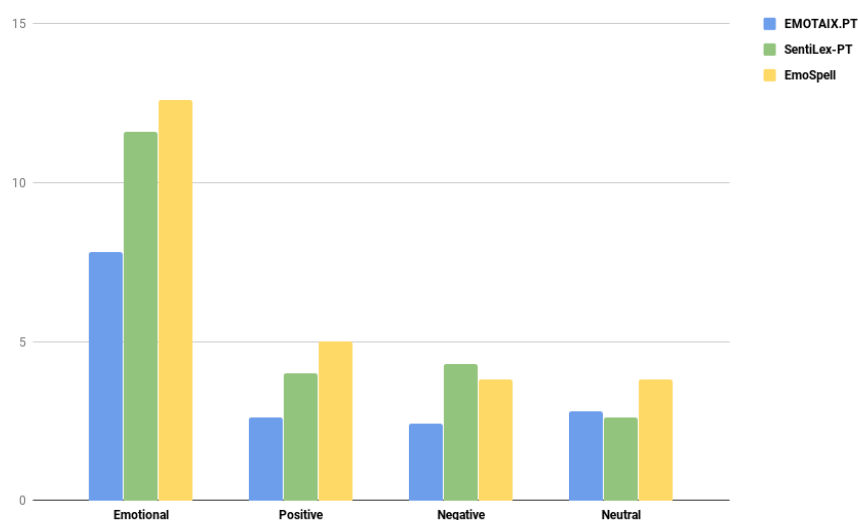


Figure 2. Bar graph representing the validation comparison.

4. Materials and Methods

4.1. Lexical Bases

The creation of EmoSpell is based on the extension of a dictionary of the morphological analyzer Jspell using the EMOTAIX.PT lexical base of emotional words. 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 also used a Sentiment Lexicon called SentiLex-PT, mostly to compare the polarities of EMOTAIX.PT and to identify relevant missing words in EMOTAIX.PT. This chapter details these three different tools used in the development of EmoSpell.

4.1.1. Jspell

The morphological analyzer Jspell (<http://natura.di.uminho.pt/wiki/doku.php?id=ferramentas:jspell>) is an extension of the Ispell spell checker. Although not a morphological analyzer itself, Ispell already includes the possibility of definition and the usage of elementary morphological rules.

Since natural language applications need to be able to handle the 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 on its origin and grammatical category. The lexical classifier is fundamental in the parsing of these types of applications.

For that purpose, Jspell uses a dictionary, defined as a list of words classified through a set of formation rules. These functionally bring an important advantage for analyzers, by simplifying the exhaustive enumeration of all the dictionary words and creating a list with the lemmas (radicals) of words and morphological rules. Consequently, flags will be associated to each word in this list [4].

The dictionary structure consists of entries, each one containing:

- **A lemma**, which is a word from where you can get others by derivation or inflection. A lemma cannot be obtained by any other lemma;
- **Morphological description**, a list of morphological properties that are key-value pairs of grammatical classification of lemmas. They may contain macros for simplification, as explained later in this section;
- **Derivation rules**, which are a set of identifiers of derivation or inflection rules (flags), defined in a separate file called affix rules.

Therefore, a typical entry in the dictionary is a line [24]:

word/classification/flags[/comment]

As can be seen, each line is composed of three or four parts, separated by division bars (/), with the fields mentioned above. For instance, the word “gato” (“cat”) has the following entry in the dictionary:

gato/CAT=nc,G=m,N=s/p/

By analysing this example, it is clear that the classification is based on the key-value pairs of characteristics and its associated categories. In Jspell, the characteristics are designated as “features” and they are as follows:

CAT—Category
G—Gender
N—Number
P—Person
T—Tense
TR—Transitivity
GR—Degree
FSEM—Semantic Role

Thus, taking the example of the word “gato”, its grammatical category, “CAT”, is “nc”, an abstract noun; its gender is masculine, and its number is singular. The last “p” is the flag, meaning that the word belongs to the rule “p”, implying that it can form the plural form.

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

Jspell uses macros in the classification field of a dictionary entry to replace enumerations of key-value pairs. The implementation of macros simplifies the repetitive classifications of all words, defining abbreviations for the macro and associating them with the classification pairs.

Macros begin with a sharp (“#”) and are represented by the abbreviation of the macro name and its definition. For example, for the transitive verbs, the macro “#vt” was defined:

#vt/CAT=v,TR=t

Thereby, a transitive verb, such as the verb “acalmar” (“soothe”), has the entry:

acalmar/#vt/XYPLM/

which is equivalent to:

acalmar/CAT=v,TR=t/XYPLM/

To conclude, Jspell can be used with several purposes, ranging from an interactive Web application with menus and options to a library. It also includes a line interpreter, where the user can write a word and receive the corresponding information. This interpreter can function with other programs interacting with Jspell through pipes. Furthermore, Jspell can be used as a standard library (dll/so/dylib), which can be an advantage in efficiency performance.

4.1.2. SentiLex-PT

As the name suggests, SentiLex-PT (<http://dmir.inesc-id.pt/project/SentiLex-PT/02>) is a sentiment lexicon, which was designed to analyze the sentiment and opinion about human entities in texts written in Portuguese. 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, in order to classify a word based on its polarity, the target of the sentiment is verified in order to identify whether it has a subject or complement function. For example, the word “fat”—a modifier of a name of human nature, (e.g., “fat guy”)—has a negative polarity, but it can have a positive one if combined with a name such as salary (e.g., “fat salary”) [6].

SentiLex-PT includes two dictionaries: *SentiLex-lem*, which describes the lemmas, and the *SentiLex-flex*, the dictionary that corresponds to the inflected forms. In the SentiLex-lem, each line includes [25]:

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

For instance, one entry of *SentiLex-lem* is:

```
enganar .PoS=V;TG=HUM:NO:N1;POL:NO=-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*, it also contains information about inflection such as gender and number.

Some examples of entries with the correspondent lemma “*bonito*” (“beautiful”) are:

```
bonita, bonito .PoS=Adj;FLEX=fs;TG=HUM:NO;POL:NO=1;ANOT=MAN
bonitas, bonito .PoS=Adj;FLEX=fp;TG=HUM:NO;POL:NO=1;ANOT=MAN
bonito, bonito .PoS=Adj;FLEX=ms;TG=HUM:NO;POL:NO=1;ANOT=MAN
bonitos, bonito .PoS=Adj;FLEX=mp;TG=HUM:NO;POL:NO=1;ANOT=MAN
```

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. EmoSpell will allow a much more detailed sentiment analysis, as each word will not only be classified based on three polarities, but by categories of sentiment as well.

4.1.3. 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) [20]. In 2009, Piolat and Bannour created a similar tool—EMOTAIX [21]—for the French lexicon. 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—EMOTAIX.PT [5].

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 by positive and negative. Negative emotions are divided into 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 composed of three additional categories of free valence: *Surprise*, *Indifference*, and *Non-Specified*. For a better understanding, Figure 3 graphically represents the different levels of organization:

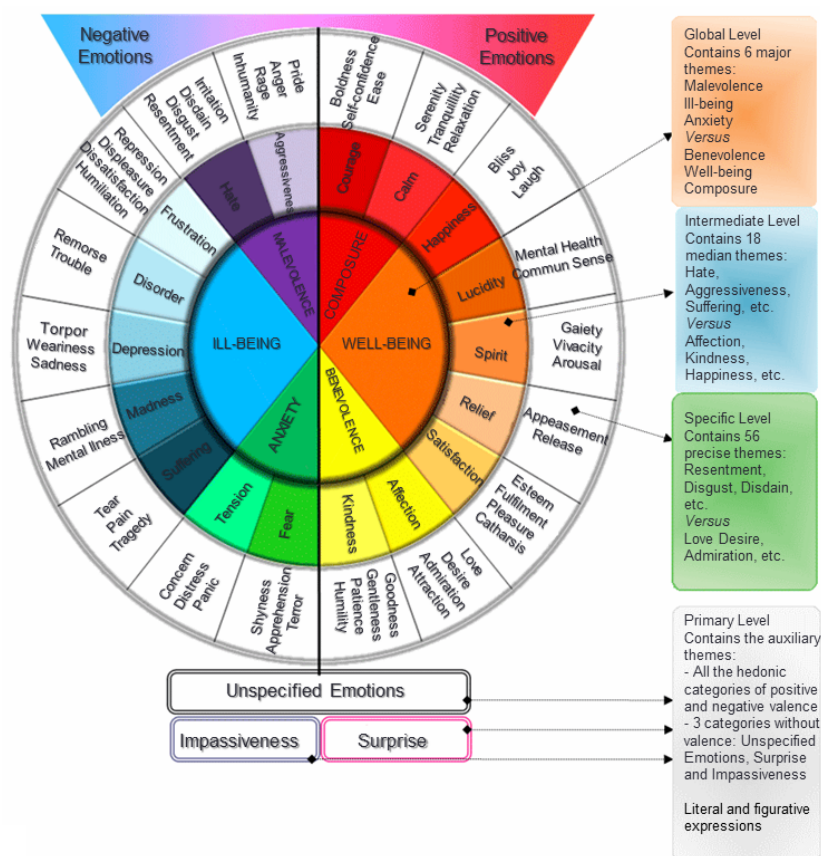


Figure 3. EMOTAIX.PT—Levels of organization (source: [5]).

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

4.2. Creation of EmoSpell

The morphological and emotional analyzer created in this project is an extension of the Jspell dictionary with the emotional annotations, accessible via two interfaces for word and text analysis.

This subsection explains the dictionary generation procedure, presents the design of the generator, and illustrates the kind of analysis provided by EmoSpell.

4.2.1. 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 multiple sources were verified for common features. In particular, EMOTAIX.PT and SentiLex-PT were checked against each other looking for inversions in polarity. Furthermore, words in SentiLex-PT missing in EMOTAIX.PT are reported for future inclusion in this lexical base.

EmoSpell 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, 52 new macros were created 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 (Unified Modeling Language) activity diagram of the Figure 4.

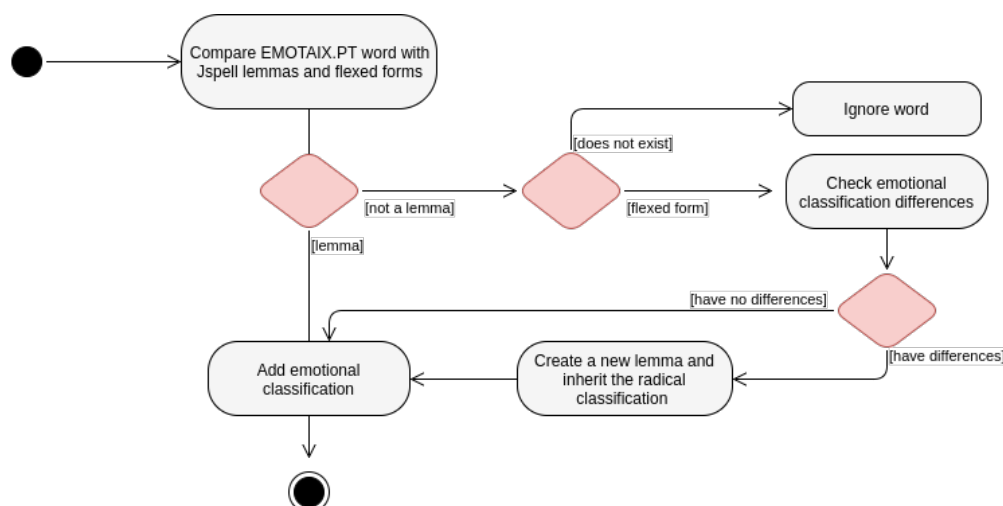


Figure 4. Dictionary generation—UML activity diagram.

All EMOTAIX.PT words were compared with Jspell’s, both lemmas and 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 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 must be added and a new entry 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 entry for the generated word. The emotional classification of the radical is “*dread*”, but for the generated word “*terrorism*” it is “*cruelty*”.

For these flexed forms, instead of creating a new entry, 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 there were only words classified as polarities. Since there are 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. These two words being associated, their classifications are the same. So, if the emotional classification of “*happy*” already existed, “*unhappy*” would also inherit this emotional value, which is not true. This problem was solved by adding a new entry for the opposite word in the dictionary. The new entry 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 each verbal form, since it is not possible to flex them from the radical, by definition. If the word from

EMOTAIX.PT is a verbal form of an irregular verb, it is necessary to add the emotional classification to the other forms.

Given that the file to be used by 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 their emotional classification.

4.2.2. Dictionary Generator

The UML diagram in Figure 5 depicts the design of the dictionary generator that implements the procedure outlined in the previous subsection. The ASGenerator class—where the EmoSpell dictionary is generated—is associated with the AnalyzerSource class. Its instances store in memory each of the three dictionary sources upon which the EmoSpell dictionary is based, as well as the EmoSpell dictionary that is generated.

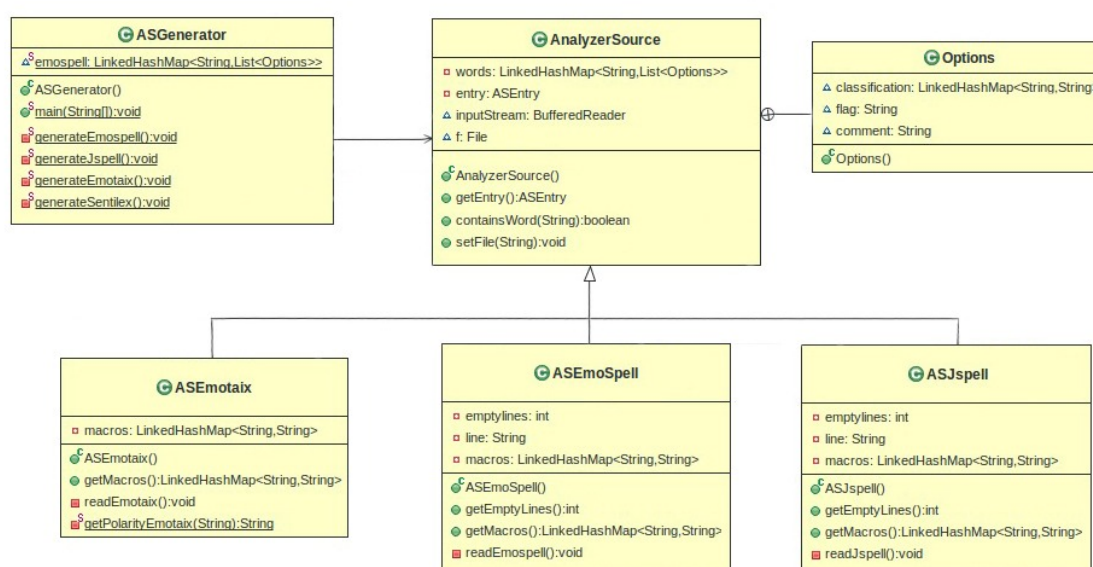


Figure 5. UML diagram of the EmoSpell dictionary generator.

For each dictionary entry the AnalyzerSource keeps a list of Options—one instance for each meaning of the word. Simultaneously, each Options instance holds the syntactic and semantic classifications assigned to a particular word meaning. Each word is associated to a list of Options, with each option being the classification for the lemma. If the word has two different classifications, two options will be available. As an example, the word “*sério*” (“serious”) has as emotional classification both “non-specific emotion” and “humility”. The ASJspell, ASSentilex, and ASEmotaix are specializations of AnalyzerSource. These classes represent the different dictionary sources and contain the specific methods to read the respective dictionary file. The specializations of AnalyzerSource can also write back their content in the format of the respective analyzer.

The method generateEmoSpell(), implemented in the ASGenerator, verifies the emotional words of the EMOTAIX.PT instance of AnalyzerSource with the words from the Jspell instance, and creates a new classification of the emotional value of words, as a result expanding the Jspell classification. The final step is to serialize the EmoSpell dictionary in the Jspell format.

4.2.3. Text Analysis

The extended dictionary enables EmoSpell to analyze words and texts with both syntactic and emotional categories. This section exposes an example of the output of the 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 the 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 (i.e., the word to be analyzed). The first three classifications are from EmoSpell. These (“CAT=nc,G=m,N=s”) are based on the morphological description of the word and, in this example, “medo” is a common name, male gender, and singular name. The following three classifications are the emotional category levels: “EmoGlobal=ansiedade” corresponds to the global emotional category (anxiety); “EmoIntermediate=medo” is the intermediate level (fear); and “EmoSpecific=pavor” refers to the most specific level, “pavor” (dread).

4.3. Interfaces

Although accessible from the command line as a Jspell dictionary, 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 the end user via a Web browser and the API enables the integration with remote applications using Web services.

4.3.1. EmoSpell GUI

The EmoSpell Web application was developed with the GWT (Google Web Toolkit). Client server communication relies on the GWT RPC (Remote Procedure Calls) framework. Figure 6 is a screenshot of the Web application. It contains a rich text area where users can write the text they want to analyze. The submitted text is classified as a “bag of words”, doing the linguistic and emotional analysis of each word. Additionally, the text is then annotated in-place with the result of word-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 3. This provides a simple and immediate understanding of the text emotional load. Other information, such as detailed emotional and syntactic categories, is displayed as a tool tip when the mouse cursor hovers over each word.

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

The GWT (Google Web Toolkit) application can be represented by the diagram in Figure 7.

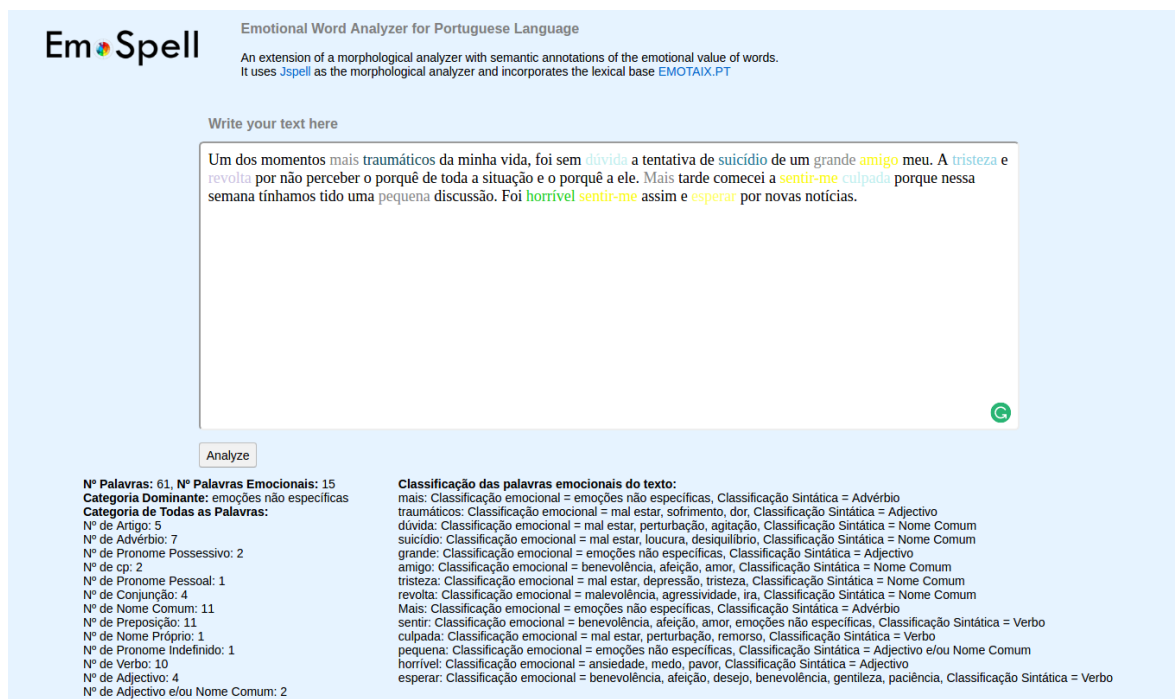


Figure 6. Graphical user interface (GUI) screenshot—analysed text with classification information.

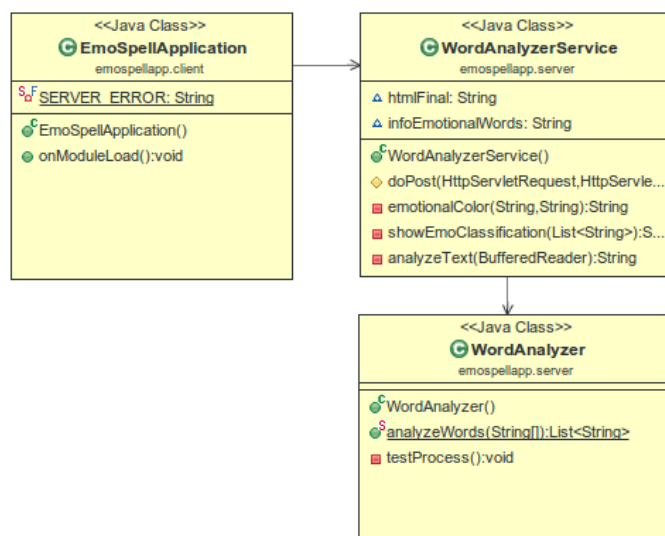


Figure 7. UML diagram for the generation of EmoSpell file.

4.3.2. EmoSpell API

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 RESTful architectural model, with a single function for analysing texts. The request is implemented as a HTTP POST method that receives the text as HTTP parameter. The response is a XML document with EmoSpell’s analysis.

The document type of the response is built on EmotionML (<https://www.w3.org/TR/emotionml/>) (Emotion Markup Language) and TEI (<http://www.tei-c.org/index.xml>) (Text Encoding Initiative) [26]. EmotionML is a W3C recommendation to represent emotions [27]. This markup language consists of a root document, with <emotionml> annotation, containing one or more <emotion> elements that represent the emotional classification and that can have more elements such as category,

action-tendency, and dimension. EmotionML fragments can also be embedded 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. These include 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.

5. Conclusions

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 governments. Nevertheless, the motivation for this research work came from the application of sentiment analysis to the research on cognitive processes in writing. Meanwhile, EmoSpell is already being integrated with HandSpy.

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 by discovering their polarity or valence. To achieve this, EmoSpell uses EMOTAIX.PT, a lexical base structured in several hierarchical levels of emotional value.

The proposed system was built on the lexical analyzer Jspell to enhance the recognition power of EMOTAIX.PT. The main contribution of this work is a procedure for the generation of a new Jspell dictionary integrating EMOTAIX.PT. A secondary contribution is the addition of 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 is 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 still challenges to overcome, such as multiple word analysis and the differentiation of the polarity.

Multiple word analysis would be an important feature for this project. The issue with the usage 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 [28]. 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.

Like 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. This means that it would solve the problem of a word possibly having opposite polarity when combined with another word.

Acknowledgments: The authors wish to thank to Alberto Simões and José João Almeida, the authors of Jspell, for their help. This work is financed by BIAL, through 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".

Author Contributions: Maria Inês Maia was responsible for the system implementation, experimental validation and results analysis. She also wrote the manuscript, in consultation with José Paulo Leal, who was responsible for the design and direction of the project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wilson, T.; Wiebe, J.; Hoffmann, P. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005; pp. 347–354.

2. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **2011**, *37*, 267–307.
3. Maia, M.I.; Leal, J.P. An emotional word analyzer for Portuguese. In *Open Access Series in Informatics, Proceedings of the 6th Symposium on Languages, Applications and Technologies (SLATE 2017), Porto, Portugal, 26–27 June 2017*; Schloss Dagstuhl – Leibniz-Zentrum für Informatik: Dagstuhl Publishing, Germany, 2017; pp. 17:1–17:14.
4. Almeida, J.J.; Pinto, U. Jspell—um módulo para análise léxica genérica de linguagem natural. Presented at Actas do X Encontro da Associação Portuguesa de Linguística, Évora, Portugal, 1994; pp. 1–15.
5. Costa, S.F.O. *Adaptação e teste de uma base lexical de palavras emocionais para o português europeu: (EMOTAIX. PT)*; Faculdade de Psicologia e de Ciências da Educação é: Porto, Portugal, 2012. (In Portuguese).
6. Silva, M.J.; Carvalho, P.; Sarmiento, L. Building a sentiment lexicon for social judgement mining. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language, Coimbra, Portugal, 17–20 April 2012*; pp. 218–228.
7. Deane, P.; Odendahl, N.; Quinlan, T.; Fowles, M.; Welsh, C.; Bivens-Tatum, J. Cognitive models of writing: Writing proficiency as a complex integrated skill. *ETS Res. Rep. Ser.* **2008**, *2008*, i-36.
8. Guinet, E.; Kandel, S. Ductus: A software package for the study of handwriting production. *Behav. Res. Methods* **2010**, *42*, 326–332.
9. Monteiro, C.; Leal, J.P. Managing experiments on cognitive processes in writing with HandSpy. *Comput. Sci. Inf. Syst.* **2013**, *10*, 1747–1773.
10. Chee, Y.M.; Froumentin, M.; Watt, S.M. *Ink markup language (InkML)*; W3C Working Draft; World Wide Web Consortium: Cambridge, MA, USA, 2006; Volume 23.
11. Bray, T.; Paoli, J.; Sperberg-McQueen, C.M.; Maler, E.; Yergeau, F. Extensible markup language (XML). *World Wide Web J.* **1997**, *2*, 27–66.
12. Liu, B. *Sentiment Analysis and Opinion Mining*; Synthesis lectures on human language technologies; Morgan & Claypool Publishers: San Rafael, CA, USA, 2012; Volume 5, pp. 1–167.
13. Chamlerwat, W.; Bhattarakosol, P.; Rungkasiri, T.; Haruechaiyasak, C. Discovering Consumer Insight from Twitter via Sentiment Analysis. *J. Universal Comput. Sci.* **2012**, *18*, 973–992.
14. O'Connor, B.; Balasubramanyan, R.; Routledge, B.R.; Smith, N.A. From tweets to polls: Linking text sentiment to public opinion time series. *Int. Conf. Web Soc. Med.* **2010**, *11*, 1–2.
15. Leung, C.W.; Chan, S.C.; Chung, F.I. Integrating collaborative filtering and sentiment analysis: A rating inference approach. In *Proceedings of the ECAI 2006 workshop on recommender systems, Riva del Garda, Italy, 28–29 August 2006*; pp. 62–66.
16. Vinodhini, G.; Chandrasekaran, R. Sentiment analysis and opinion mining: A survey. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2012**, *2*, 282–292.
17. Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; Huang, R. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013*.
18. Mohammad, S.M. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), San Diego, CA, USA, 12–17 June 2016*; pp. 174–179.
19. Stone, P.J.; Dunphy, D.C.; Smith, M.S. *The General Inquirer: A Computer Approach to Content Analysis*; Cambridge: London, UK, 1966.
20. Pennebaker, J.W.; Francis, M.E.; Booth, R.J. Linguistic inquiry and word count: LIWC 2001. In *Mahway: Lawrence Erlbaum Associates; LIWC.net*: Austin, TX, USA, 2001.
21. Piolat, A.; Bannour, R. EMOTAIX: un scénario de Tropes pour l'identification automatisée du lexique émotionnel et affectif. *L'Année Psychologique* **2009**, *109*, 655–698. (In French).
22. Fellbaum, C. *WordNet*; Wiley Online Library: Hoboken, NJ, USA, 1998.
23. Esuli, A.; Sebastiani, F. SENTIWORDNET: A high-coverage lexical resource for opinion mining. *Evaluation* **2007**, 1–26.
24. Simões, A.; Almeida, J.J. *jspell. pm: um módulo de análise morfológica para uso em processamento de linguagem natural*; Associação Portuguesa de Linguística (APL): Évora, Portugal, 2001. (In Portuguese).
25. Carvalho, P.; Silva, M.J. SentiLex-PT: Principais características e potencialidades. *Oslo Stud. Lang.* **2015**, *7*, 1. (In Portuguese).

26. Ide, N.; Véronis, J. *Text Encoding Initiative: Background and Contexts*; Springer Science & Business Media: Berlin, Germany, 1995; Volume 29.
27. Burkhardt, F.; Pelachaud, C.; Schuller, B.W.; Zovato, E.; Emotion, M.L. *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*; Dahl, D.A., Ed.; Springer: Cham, Switzerland, 2017; pp. 65–80.
28. Asmi, A.; Ishaya, T. Negation identification and calculation in sentiment analysis. In Proceedings of the Second International Conference on Advances in Information Mining and Management, Venice, Italy, 21–26 October 2012; pp. 1–7.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Copyright of Information (2078-2489) is the property of MDPI Publishing and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.