Lingmotif: Sentiment Analysis for the Digital Humanities

Antonio Moreno-Ortiz
University of Málaga
Spain
amo@uma.es

Abstract

Lingmotif is a lexicon-based, linguistically-motivated, user-friendly, GUI-enabled, multi-platform, Sentiment Analysis desktop application. Lingmotif can perform SA on any type of input texts, regardless of their length and topic. The analysis is based on the identification of sentiment-laden words and phrases contained in the application's rich core lexicons, and employs context rules to account for sentiment shifters. It offers easy-to-interpret visual representations of quantitative data (text polarity, sentiment intensity, sentiment profile), as well as a detailed, qualitative analysis of the text in terms of its sentiment. Lingmotif can also take user-provided plugin lexicons in order to account for domain-specific sentiment expression. Lingmotif currently analyzes English and Spanish texts.

1 Introduction

Lingmotif[1] is a lexicon-based Sentiment Analysis (SA) system that employs a set of lexical sources and analyzes context, by means of sentiment shifters in order to identify sentiment-laden text segments and produce two scores that qualify a text from a SA perspective. In a nutshell, it breaks down a text into its constituent sentences, where sentiment-carrying words and phrases are searched for, identified, and assigned a valence (i.e., a sentiment index). The overall score for a text is computed as a function of the accumulated negative, positive and neutral scores. Specific domains can be accounted for by applying

user-provided dictionaries, which can be imported from CSV files, and used along with the application's core dictionary.

Lingmotif's SA approach could be loosely characterized as bag-of-words, since sentiment is computed solely based on the presence of certain lexical items. However, Lingmotif is not just a classifier. It also offers a visual representation of the *sentiment profile* of texts, allows to compare the profile of multiple documents side by side, and can process ordered document series. Such features are useful in discourse analysis tasks where sentiment changes are relevant, whether within or across texts, such as political speeches and narratives, or to track the evolution in sentiment towards a given topic (in news, for example).

Being focused on the end user, Lingmotif uses a simple, easy-to-use GUI that allows users to select input and options, and launch the analysis (see Figure 1). Results are generated as an HTML/Javascript document, which is saved to a predefined location and automatically sent to the user's default browser for immediate display. Internally, the application generates results as an XML document containing all the relevant data; this XML document is then parsed against one of several available XSL templates, and transformed into the final HTML.

Lingmotif is available for the Mac OS, MS Windows, and Linux platforms. It is free for non-commercial purposes.

2 Lexicon-based Sentiment Analysis

Lingmotif is a lexicon-based SA system, since it uses a rich set of lexical sources and analyzes context in order to identify sentiment laden text segments and produce two scores that qualify a text from a SA perspective. In a nutshell, it breaks down a text into its constituent sentences,
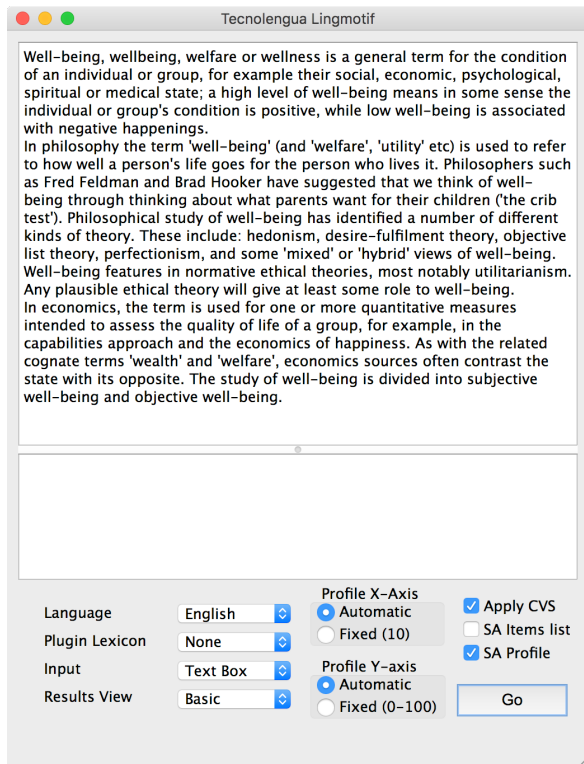

Figure 1: Lingmotifs GUI

where sentiment-carrying words and phrases are searched for, identified, and assigned a valence. For each language, Lingmotif uses a core lexicon, a set of context rules, and, optionally, one or more plugin lexicons.

In the following sections, the most salient aspects of the application's sentiment analysis engine are described. A more thorough description can be found in (Moreno-Ortiz, 2017).

2.1 Core sentiment lexicon

A lexical item in a Lingmotif lexicon can be either a single word or a multiword expression. Each entry is defined by a specification of its form, part of speech, and valence. The valence is an integer from -5 to -2 for negatives and 5 to 2 for positives. The items form can either be a literal string or a lemma. For the part-of-speech specification, Lingmotif uses the Penn Treebank tag set. A wildcard (ALL) can be used for cases where all possible parts of speech for that lemma share the same valence. Sentiment disambiguation is currently dealt with using exclusively formal features: part-of speech tags and multi-word-expressions. MWEs usually include words that may or may not have the same polarity of the expression. including such expressions can solve disambiguation for

many cases. For example, we can classify as negative the word kill and then include phrases such as kill time with a neutral valence. When this is not possible, the options are to include it with the more statistically probable polarity or simply leave it out when the chances of getting the item with one polarity or another are similar.

2.2 Context rules

Context rules are Lingmotifs mechanism to deal with sentiment shifters. They work by specifying words or phrases that can appear in the immediate vicinity of the identified sentiment word. Basically, we use the same approach as (Polanyi and Zaenen, 2006). Previous implemented systems following this approach are (Kennedy and Inkpen, 2006), (Taboada et al., 2011), and (Moreno-Ortiz et al., 2010). We use simple addition or subtraction (of integers on a -5 to 5 scale in our case). When a context rule is matched, the resulting text segment is marked as a single unit and assigned the calculated valence, as specified in the rule. Lingmotifs context rules were compiled by extensive corpus analysis, studying concordances of common polarity words (adjectives, verbs, nouns, and adverbs), and then testing the rules against texts to further improve and refine them.

2.3 Plugin lexicons

Topic has been consistently shown to determine the semantic orientation of a text (Aue and Gamon, 2005), (Pang and Lee, 2008). Being a general-purpose SA system, Lingmotif provides a flexible mechanism to adapt to specific domains by means of user-provided lexicons. Lexical information contained in plugin lexicons overrides Lingmotifs core lexicon, providing domain-specific sentiment items. They can be created as a CSV file following a simple format, which is then imported into Lingmotif's internal database.

3 Single and multi-document modes

From a classification perspective, it only makes sense to use a large set of texts to be analyzed (i.e., classified). However, since Lingmotif is able to specifically identify and mark those text segments that convey sentiment, we can take advantage of this feature to measure sentiment not only in the text as a whole, but in subsections of the text, producing a *sentiment map* of the text and display the result in several ways.

A single input text can be typed or pasted in the text box area, or text files can be loaded, depending on the selected input type. Loading files allows the user to select one of them and analyze it in single-mode, or select the complete set of files.

3.1 Single-document mode

For every text analyzed, either in single or multi-document mode, Lingmotif produces a number of metrics for each individual text. The two metrics that summarize the text's overall sentiment are the Text Sentiment Score (TSS) and the Text Sentiment Intensity (TSI). Both are displayed by means of visual, animated (Javascript) gauges at the top of the results page. The numeric indexes (on a 0-100 scale) are categorized in ranges, from "extremely negative" to "extremely positive", to make numeric results more intuitively interpretable by the user. Both gauges are also color and intensity-coded in the red (negative) to green (positive) range (see Figure 2).
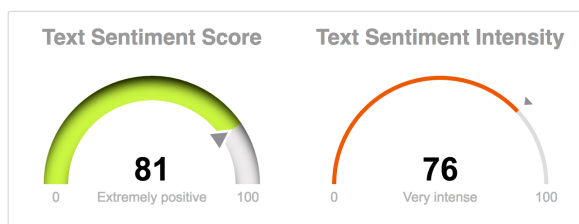

Figure 2: Sentiment scores gauges

For long texts, Lingmotif will also generate a *sentiment profile*, which is a visual representation of the text's internal structure and organization in terms of sentiment expression. This Javascript graph is interactive: hovering the data points will display the lexical items that make up that particular text segment (see Figure 3).
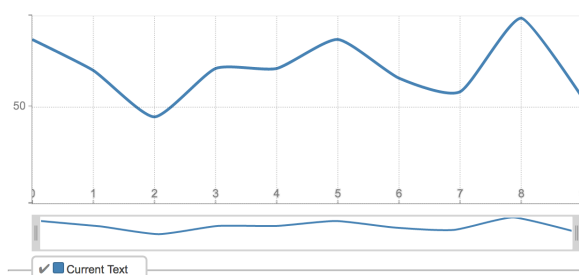

Figure 3: Document sentiment profile

The next three sections of the results document are shown in Figure 4 below. First is the quantitative data tables, which include common text metrics and a breakdown of the sentiment analysis data for the analyzed text. The final results section shows the input text after processing, where the identified sentiment items are color-coded to represent their polarity. This makes it possible to know exactly what the analyzer found in the text.

Figure 4: Sentiment scores gauges

An option exists (*advanced view*) to also show specific data for each lexical item

Figure 5: Detail in Advanced Mode

3.2 Multi-document analysis

Multiple input texts can be analyzed in one of several modes (see below). When in multi-document mode, Lingmotif will analyze documents one by one, generating one HTML file for each, although they will not be displayed on the browser, just saved to the output folder. When the analysis is finished, a single results page will be displayed. This page is a summary of results, and is different from the single-document results page: the gauges for TSS and TSI are now the average for the analyzed set and the detailed analysis section contains a quantitative analysis of each of the files in the set. The first column in this table shows the title of the document (file name without extension) as a hyperlink to the HTML file for that particular file.

Available multi-document analysis modes are the following:

- **Classifier** (default): a stacked bar graph and data table are offered showing classification

results based on their TSS category. The graph offers a visualization of results (see Figure 6); both its legend and the graph itself are interactive. A table summarizing the classification results is also offered.
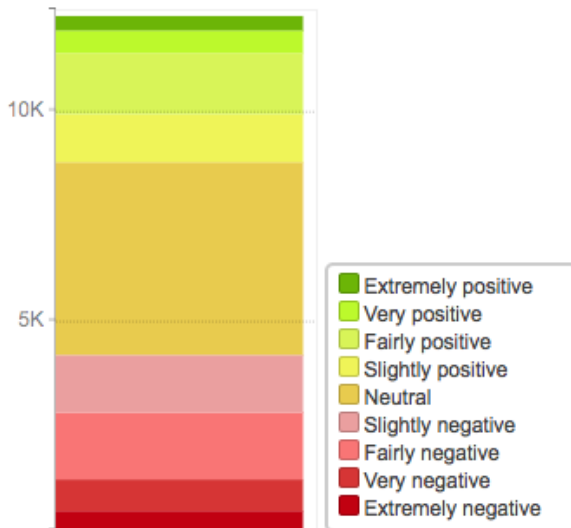

Figure 6: Classifier graph

To facilitate analysis of large sets of documents, they can be loaded from a single text file where each line is assumed to be an individual document. Lingmotif classifies documents according to their TSS, which will always include the neutral category.

- **Series**: the set of loaded files is assumed to be in order, chronological (time series) or otherwise. Each data point in the Sentiment Analysis Profile represents one document. The data point is the average TSS for that particular document.

- **Parallel**: produces a graph with one line for each file (this mode is limited to 15 documents). This is useful to compare sentiment flow in texts side by side (see Figure 7).

- **Merge**: this is a convenience option merges all loaded individual files into one single text.

4 Conclusions

Lingmotif goes beyond what SA classifiers have to offer. It offers automatic identification of sentiment-laden words and phrases, as well as text segments. Its many visual representations of the text's structure from a sentiment perspective make
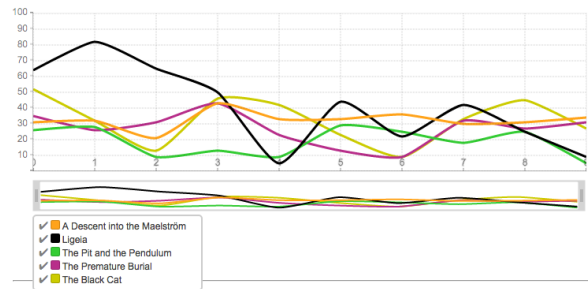

Figure 7: Multi-document analysis in parallel mode (5 documents)

it a valuable tool for research in the Digital Humanities where such tasks are relevant. The possibility to integrate user-provided lexicons enables it to adapt to any subject domain.

References

A. Aue and M. Gamon. 2005. Customizing sentiment classifiers to new domains: A case study.

A. Kennedy and D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence* 22(2):110–125.

A. Moreno-Ortiz. 2017. Lingmotif: A user-focused sentiment analysis tool. *Procesamiento del Lenguaje Natural* 58:21–29.

A. Moreno-Ortiz, Á. Pérez Pozo, and S. Torres Sánchez. 2010. Sentitext: sistema de análisis de sentimiento para el español. *Procesamiento del Lenguaje Natural* 45:297–298.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(12):1–135.

L. Polanyi and A. Zaenen. 2006. Contextual valence shifters. In *Computing Attitude and Affect in Text: Theory and Applications*, Springer, Dordrecht, The Netherlands, volume 20 of *The Information Retrieval Series*, pages 1–10. Shanahan, james g., qu, yan, wiebe, janyce edition.

M. Taboada, J. Brooks, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37(2):267–307.