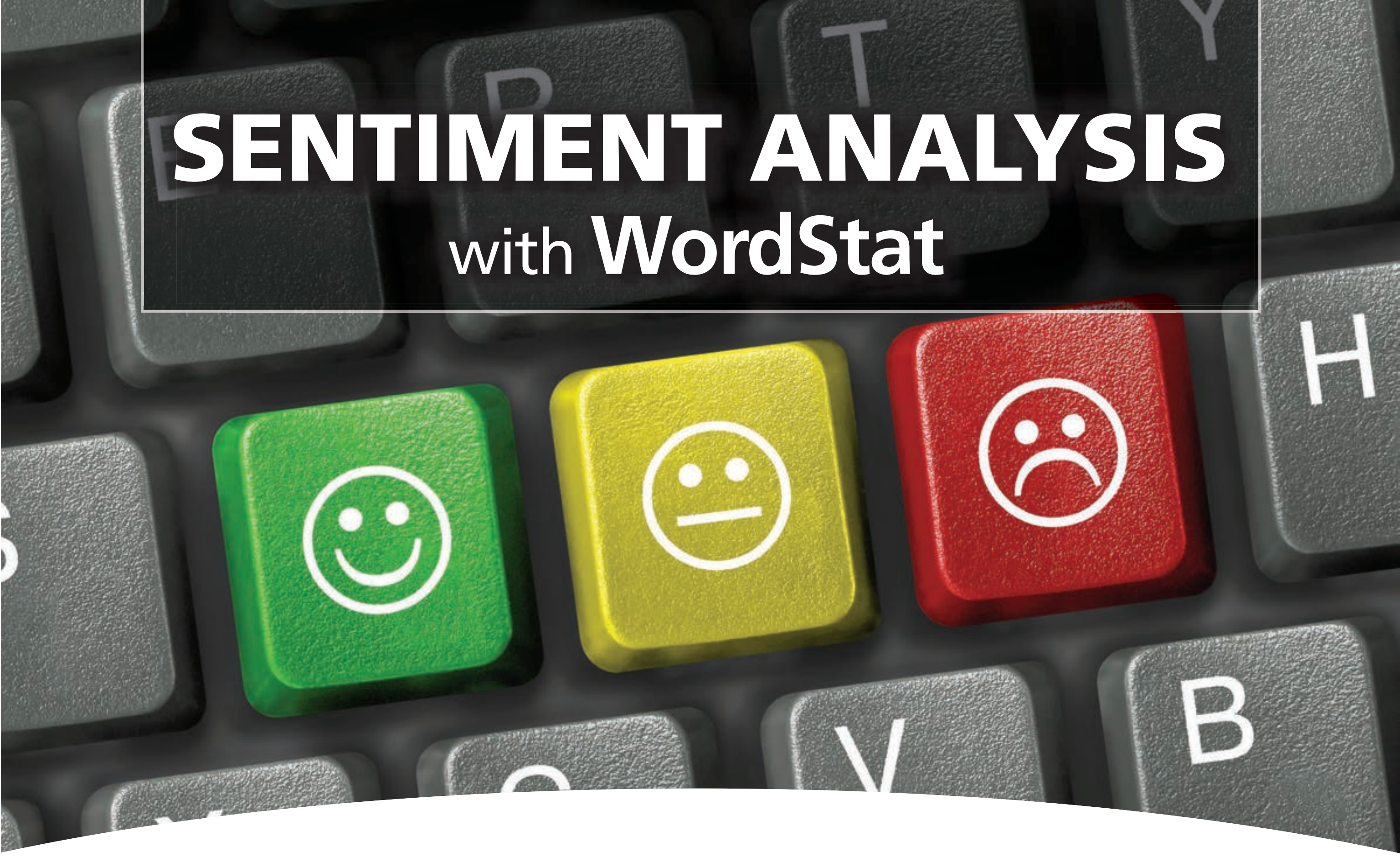
SENTIMENT ANALYSIS with WordStat



What is Sentiment Analysis?

Automated sentiment analysis is an application of text analytics techniques for the identification of subjective opinions in text data. It normally involves the classification of text into categories such as "positive", "negative" and in some cases "neutral". Over the last five years, we have seen a tremendous increase in demand for sentiment analysis tools by companies collecting people's opinions of the company and its products and services but also by social science researchers. To fulfill the increasing demands these kinds of tools, more and more researchers and companies are releasing products to perform sentiment analysis. Many of them claim to be able to perform sentiment analysis of any type of document in every domain. Unfortunately, experience has shown us that, an "out-of-the-box" sentiment analysis tool working across domains does not yet exist. The main reason sentiment analysis is so difficult is that words often take different meanings and are associated with distinct emotions depending on the domain in which they are being used. The use of a word like "fingerprints" may represent a major breakthrough in a criminal investigation but a major headache for smartphone manufacturers. There are even situations where different forms of a single word will be associated with different sentiments. For example, we found in customer feedback that the word "improved" was associated with positive comments, but"improve" was more often used in negative ones.

PROVALIS RESEARCH

All sentiment analysis tools rely, to varying degrees, on lists of words and phrases with positive and negative connotations or are empirically related to positive or negative comments. We have used such a list in the past for sentiment analysis tasks, yet we have never made available our sentiment dictionary for several reasons. Such lists cannot be used as is, but need to be customized to specific domains in order to provide reliable results. A lot of effort is needed to develop a domain-specific sentiment dictionary and to identify the proper vocabulary associated with the expression of positive and negative feelings. Many people are not necessarily willing to spend time performing such customization and validation tasks. They want something that they believe will work right away and they would be ready to pay a lot for such a tool.

We believe there is a risk some people may use our sentiment dictionary as is, without attempting to validate it or customize it to their own type of data. Those who are aware of the limitations of such lists may still have no idea how such customization could be achieved and need some guidance. However, despite the potential misuse of sentiment analysis word lists, we have decided to make our WordStat Sentiment Dictionary available to the public. One of the reasons that made us change our minds was the publication of two articles. The first of these, written by Loughran and McDonald (2011), stresses the danger of using dictionaries like ours without any attempt to adapt them to the intended domain, in their case accounting and financial news. The researchers developed their own domain-specific sentiment dictionaries and describe, in some detail, the process by which they selected words and validated their results. The second paper, published by Young and Soroka (2011), also presents the construction and validation process of a sentiment dictionary but this time customized for the analysis of political news. Both papers represent commendable efforts and are worth reading by anyone who would like to learn how to create a context-specific sentiment analysis dictionary.

The Loughran and McDonald Financial Sentiment Dictionary

The Loughran and McDonald (2011) article provides a clear demonstration that applying a general sentiment word list to accounting and finance topics can lead to a high rate of misclassification. They found that about three-fourths of the negative words in the Harvard IV TagNeg dictionary of negative words are typically not negative in a financial context. For example, words like "mine", "cancer", "tire" or "capital" are often used to refer to a specific industry segment. These words are not predictive of the tone of documents or of financial news and simply add noise to the measurement of sentiment and attenuate its predictive value. The authors created custom lists of negative and positive words specific to the accounting and financial domain. Another benefit of the dictionary they propose is that it shows how quantitative content analysis can move beyond mere dichotomous differentiations typical of sentiment analysis and can also be used to measure additional dimensions of interest. Two noteworthy additions are the Uncertainty word list that attempts to measure the general notion of imprecision (without an explicit reference to risks), and the Litigiousness word list that may be used to identify potential legal problem situations. They also included Weak Modal and Strong Modal word lists. The following table illustrates the various categories of the Loughran and McDonald Financial Sentiment Dictionary:

LOUGHRAN AND MCDONALD FINANCIAL SENTIMENT DICTIONARY

SCALE	NO. OF WORDS	SAMPLE WORDS
Negative	2,337	termination, discontinued, penalties, misconduct, serious, noncompliance, deterioration, felony
Positive	353	achieve, attain, efficient, improve, profitable
Uncertainty	285	approximate, contingency, depend, fluctuate, indefinite, uncertain, variability
Litigiousness	731	claimant, deposition, interlocutory, testimony, tort
Weak Modal Words	27	could, depending, might, possibly
Strong Modal Words	19	always, highest, must, will

Lexicoder Sentiment Dictionary (LSD)

Young and Soroka (2011) explain the construction process of their sentiment analysis dictionary. Their objective was to expand the score of coverage of existing sentiment dictionaries, without compromising accuracy. Like the article cited previously, it relies partly on the Harvard IV dictionary (Stone et al., 1966), but has added to those initial positive and negative words, other entries from Roget's Thesaurus as well as from Colin Martindale's Regressive Imagery Dictionary (both dictionaries available in WordStat format). They removed neutral and ambiguous words and then extracted the most frequent ones, resulting in a list of 2,858 negative entries and 1,709 positive ones. Some noteworthy features of their dictionary are the implementation of basic word sense disambiguation with the use of phrases, truncation and preprocessing, as well as the effort to deal with negations. To assess the accuracy of the LSD, the dictionary was tested against a body of 900 human-coded news stories. Results suggest that their dictionary is more highly related to ratings by human coders than other available dictionaries. The authors also established the predictive validity of the dictionary by demonstrating high correlations between sentiment scores and the evolution of poll results during the 2006 Canadian federal election campaign.

WordStat Sentiment Dictionary 1.2

The WordStat Sentiment Dictionary was actually designed by combining negative and positive words from the Harvard IV dictionary, the Regressive Imagery dictionary (Martindale, 2003) and the Linguistic and Word Count dictionary (Pennebaker, 2007). The WordStat dictionary building utility program was then used to expand its word list by automatically identifying potential synonyms and related words as well as any inflected forms. We ended up with more than 9164 negative and 4847 positive word patterns. Actually, sentiment is not measured with those two lists of words and word patterns but instead with two sets of rules that attempt to take into account negations that may precede those words. For example, negative sentiment is measured by using the following two rules:

- Negative words not preceded by a negation (no, not never) within three words in the same sentence.
- Positive words not preceded by a negation within three words in the same sentence.

Positive sentiment is measured in a similar way by looking for positive words not preceded by a negation as well as negative terms following a negation. However, our own experiences suggest that this last rule has less predictive value and may even slightly deteriorate the measurement of sentiments. But there may be some situations where such a rule could help predict positive sentiments. We decided to keep this last rule and let the user decide whether it should be applied or not.

Recommended Use

WE DO NOT RECOMMEND USING THIS DICTIONARY AS IS. We strongly believe that doing so will not present accurate results. We recommend instead customizing this dictionary by applying the following procedures:

1. Remove Domain-Specific Words - Identify and remove frequent words that may be specific to your domain of interest and that usually do not have positive or negative connotations. Reviewing all of those words may be time consuming, so a more time-efficient way to do this would be to apply this dictionary to a large set of documents in your domain area and identify words that appear frequently. You should then use the keyword-in-context features of WordStat to assess how those words are being used.

2. Identify Wrongful Predictions - If you have a set of documents that have already been categorized as positive or negative, or contain satisfaction scores or any other author-sentiment indicator, we suggest using the WordStat cross-tab feature to assess the correlation between frequent positive and negative words and those indicators. From such a list, pay close attention to any word that seems to be inversely related to the expected prediction. Using the keyword-in-context feature, examine how those words are being used. If they are usually preceded by a negation (within three words), you can keep those words in the dictionary since WordStat contains rules that will take those into account.

3. Add Domain-Specific Sentiment Words and Phrases - Quite often there are specific words in your domain area that are used to refer to positive or negative aspects or features. For example, if you sell smartphones, items like "fingerprint," "noise", "drop" or "sound quality" may be highly associated with positive or negative feedback. For car manufacturers, "blind spot" "hard plastic" "chug" "whiplash" "bouncing" or any mention of "wind" or "legs" may also be related to specific opinions about a specific car. If you have access to a collection of positive and negative evaluations, one easy way to identify those domain-specific words would be to correlate the most frequent words with satisfaction scores and identify those that are highly predictive of negative and positive scores. There is, however, a trap to avoid when selecting those predictors based on their high correlation to satisfaction scores: The obtained sentiment measure may become insensitive to changes. For example, if many people complain about the poor sound quality of a cell phone, then the phrase "sound quality" will likely be highly predictive of negative comments. If in reaction to those evaluations the manufacturer releases a new version with improved sound quality, then any new positive comments about this improved sound quality may be wrongly classified as negative. This lack of sensitivity to changes is also a pitfall of many machine-learning approaches to sentiment analysis.

If you believe any word or phrase is missing or if you identify any error that should be fixed to improve the dictionary's accuracy, please let us know. Also, if you have developed any customized version of this dictionary, we would very much like to know about your efforts.

Obtaining our Sentiment Analysis Dictionaries

All three sentiment analysis dictionaries are available for free. Instructions on how to obtain any one of these dictionaries is available from our web site at:

Provalisresearch.com/sentiment

References

Loughran, T. & McDonald, B. (2011)*When is a liability not a liability?* Textual Analysis, Dictionaries and 10-Ks. *The Journal of Finance*, *66*(1), 35-66.

Young, L. & Soroka, S. (April2011). *Affectivenews: The automated coding of sentiment in political texts*, forthcoming in Political Communication.

Provalis Research (2012). *WordStat Sentiment Dictionary*, v1.2. Montreal: Provalis Research.