Jan Kochanowski University Press

This is a contribution from *Token*: *A Journal of English Linguistics*
Volume 13/2021.
Edited by John G. Newman, Marina Dossena and Sylwester Łodej.

Comparison of key statistical instruments used in lexicon-based tools for sentiment analysis in the English language

Łukasz Stolarski
Jan Kochanowski University of Kielce

ABSTRACT

This study investigates "key statistical instruments", such as the mean or the sum, used in obtaining numeric polarity scores in lexicon-based tools for sentiment analysis. First, a large number of texts rated for sentiment intensity by independent human judges was collected. Next, 15 different sentiment lexicons were used to generate sets of numeric values for each of the texts. Then, the key statistical instruments were calculated on the basis of these results and compared with the corresponding human scoring using tests for association between paired samples. The results of these tests were further examined with the use of ANOVA and Tukey HSD post-hoc analysis. The broad conclusion drawn from the analysis is that the mean, all other things being equal, is the most reliable key statistical instrument for obtaining numeric polarity scores that are similar to scores provided by human assessors. These results may be of particular importance for both developers of lexicon-based programs performing sentiment analysis and users of such software packages.

Keywords: sentiment analysis, opinion mining, sentiment mining, opinion extraction.

1. Background

Sentiment analysis is defined as "the polarity of an opinion item which either can be positive, neutral or negative" (Borth et al. 2013: 223) or a procedure which involves "determining the evaluative nature of a piece of text" (Kiritchenko et al. 2014: 723). More specifically, the term is typically used in reference to "an active area of study in the field of natural language processing that analyses people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in

DOI:10.25951/4827

text" (Hutto – Gilbert 2014: 217) or, in short, to "the computational treatment of opinion, sentiment, and subjectivity in text" (Pang – Lee 2008: 10). The subject has been explored in a large number of publications, and several reviews of literature on sentiment analysis are available (e.g. Liu 2012; Liu – Zhang 2012; Pang – Lee 2008).

Tools for performing sentence-level sentiment analysis are frequently divided into two major categories. The first one, which may be referred to as "the machine learning approach" (Ribeiro et al. 2016; Taboada et al. 2011), involves labelled training data which are used for building a classifier. Such tools are used for conducting sentiment analysis on particular types of texts, as they usually perform very well in the domain that they were trained on. Nevertheless, their performance may drop considerably in other domains (Aue – Gamon 2005) and they do not cope well with effects of linguistic context such as negation or intensification (Taboada et al. 2011: 269). "The lexicon-based approach", on the other hand, makes use of a list of words, or "a sentiment lexicon", in which each word or phrase is assigned a sentiment value. Some such lexicons involve categorical classification. An example of this is the NRC Emotion Lexicon (Mohammad – Turney 2010, 2013), which contains, among other categories, the binary distinction between "positive" and "negative". Other lexicons offer continuous polarity scores, as is the case for all the lexicons described in Section 3.2. Sentiment lexicons also differ in the way they are obtained. Some are created manually, usually involving a group of participants whose task is to label selected words in terms of sentiment polarity or value. Such tasks tend to be costly, time-consuming and labour-intensive; hence, the resulting lexicons are relatively small. Typically, they contain a few thousand words. However, they tend to be less domain specific and the tools that utilize them are usually more consistent across domains (Taboada et al. 2011). Examples of lexicons which involve human annotation are the aforementioned NRC Emotion Lexicon, Sentiment Composition Lexicon for Negators, Modals, and Degree Adverbs (SCL-NMA) described in Section 3.2.6, as well as, at least partially, the lexicons used in SentiStrength and Vader projects (see Sections 3.2.2 and 3.2.5, respectively). A large number of sentiment lexicons are, nevertheless, created automatically using seed words. They tend to contain a greater number of unigrams and sometimes longer expressions (bigrams and trigrams), but their performance may be less consistent across domains. Most of the lexicons described in Section 3.2.6 were created in this way.

Several benchmark comparisons of sentiment analysis tools have recently been published (e.g. Abbasi et al. 2014; Diniz et al. 2016; Gonçalves et al. 2013; Ribeiro et al. 2016). They demonstrate that, on average, some

software packages perform better than others; however, there is no clear winner for all possible testing sets. The performance of individual sentiment analysis tools varies depending on the domain which is being investigated.

It should be noted that studies on sentiment analysis frequently discuss potential problems which may affect results. The most pressing issues include negation and intensification. These two aspects have received much attention and numerous solutions have been suggested. For example, it was initially proposed that negation could be dealt with by reversing the polarity of a lexical item (Choi – Cardie 2008; Kennedy – Inkpen 2006). This approach, however, has been shown to be fundamentally flawed (Kennedy – Inkpen 2006; Kiritchenko et al. 2014; Taboada et al. 2011); thus, alternative solutions, such as shifting the polarity by a fixed amount, have been used (Taboada et al. 2011). A useful taxonomy of problems affecting sentiment analysis is offered in Abbasi et al. (2014). It demonstrates that even though the performance of some tools may be promising, there is still much room for improvement, and further research is necessary.

2. Aims

This paper focuses on lexicon-based tools that perform sentiment analysis on phrases, sentences and longer utterances and give continuous polarity scores. When using such tools, it becomes clear that they consist of two largely independent components. The first is a sentiment lexicon, or a group of such lexicons. The second is a sentiment analysis algorithm which calculates the final score on the basis of several (modified) digits representing sentiment values of individual words or phrases. These values may simply be added, but other solutions are also possible, e.g. calculating the mean, median or obtaining the highest absolute value.

The major aim of this project is to compare the effectiveness of such key statistical instruments in calculating final sentiment scores (for the description of the exact methods tested see Section 3.1). They are necessary at the final levels of sentiment analysis performed by tools within the lexicon-based approach. Consequently, investigating the efficacy of such statistics, other things being equal, may help in improving the overall performance of sentiment analysis tools. Additionally, the results of this study may also be useful for the end users of such software packages. In some cases, the user may choose between various ways in which the final sentiment score is calculated (e.g. SentiStrength).

It must be stressed that this study is not a benchmark comparison of any software packages. Rather than testing actual sentiment analysis tools, the current analysis focuses on the efficacy of key statistical instruments applied to "bare" sentiment lexicons. The resulting correlation with human scores is expected to be lower than the corresponding correlation obtained using complete software packages for sentiment analysis. Such packages may involve various additional strategies to deal with the problems mentioned in Section 1. Nevertheless, testing key statistical instruments on "bare" sentiment lexicons is fundamental to lexicon-based sentiment analysis and, as suggested in the previous paragraph, may be essential in improving the performance of actual software packages.

3. Methods

In order to accomplish the major aim outlined in Section 2, five key statistical instruments were defined (see Section 3.1). Next, a group of sentiment lexicons with continuous polarity scores was selected (see Section 3.2). After that, a representative number of validation texts with numerical scores for the positive-negative dichotomy provided by human respondents was obtained (see Section 3.3). Then, the key statistical instruments for each validation text were calculated on the basis of each sentiment lexicon. This task involved some text preprocessing. Each case required a slightly different approach and the details on the preprocessing are provided in the description of each lexicon. Finally, statistical tests were performed on the data obtained (see Sections 3.4 and 4).

3.1 The key statistical instruments

The five key statistical instruments chosen for comparison are summarised below.

- MEAN1 – the mean obtained on the basis of all scores, excluding the value of 0.0 added to lexical items not recognized in a given lexicon or stop words removed from the analysis. In the example presented in Table 1, the mean would be calculated as follows:
  $(0.9765 + 0.7181 – 0.3638 – 1.4753) / 4 = –0.0361$.
- MEAN2 – the mean obtained on the basis of all scores, including the value of 0.0 added to lexical items not recognized in a given lexicon or

stop words removed from the analysis. The result for the example in Table 1 would be determined in the following way:
(0.9765 + 0.7181 – 0.3638 – 1.4753) / 10 = –0.0144.

- MEDIAN – the median obtained on the basis of all scores, excluding the value of 0.0 added to lexical items not recognized in a given lexicon or stop words removed from the analysis. For the example in Table 1, MEDIAN = 0.1775.
- LAV (largest absolute value) – the largest value in all the scores, regardless of the polarity. For the example in Table 1, LAV = –1.4753.
- SUM – the sum obtained from all the scores. For the example in Table 1, SUM = –0.1445.

MEAN is a measurement which could easily be applied in obtaining final scores for sentiment intensity in lexicon-based tools. It is offered as one of several options in the GUI distribution of SentiStrength (see Section 3.2.2). This statistical instrument will be investigated in the two versions described above to see if the inclusion/exclusion of elements with no sentiment scoring affects the results. MEDIAN, to the best of the author's knowledge, has not been used in sentiment analysis software packages, but is appropriate in this study. It has characteristics similar to those of MEAN, as its purpose is to summarise datasets, but it is less affected by outliers. By contrast, LAV represents only the most extreme value in a dataset. This statistical instrument is offered as one of the options in the GUI distribution of SentiStrength. Finally, SUM is probably the most obvious solution applied in the calculation of final scores. It is used, for instance, in Vader (see Section 3.2.5) and Afinn (Nielsen 2011).

In addition to the above statistics, other possible calculations were also considered. For instance, MEDIAN could also have been calculated on the basis of all scores, including the zeroes assigned to items not recognized in a lexicon or stop words. However, the results yielded 0.0 in too many cases; thus, the method was considered significantly less reliable than the other five. Additionally, "mode" was also excluded from the analysis because it is not appropriate for continuous data.

Table 1. Example results for a text containing both positive and negative lexical items

	he	is	friendly	and	funny	but	also	naive	and	irresponsible
scores	0.0	0.0	0.9765	0.0	0.7181	0.0	0.0	–0.364	0.0	–1.4753

3.2 Sentiment lexicons

Fifteen different sentiment lexicons were used in this study. All of them involve numerical scoring. Five are associated with independent projects (SenticNet, SentiStrength, SentiWordNet, UMass Amherst Linguistics Sentiment Corpora, Vader), nine belong to the set of sentiment lexicons created by the National Research Council Canada (NRC) and the last one was created on the basis of these nine lexicons (see the last paragraph on "NRC Combined" in Section 3.2.6). All the lexicons are described in Sections 3.2.1 to 3.2.6. In each case, a general summary is presented and the way a given lexicon was used in the present study is summarized.

3.2.1 SenticNet

SenticNet (Cambria et al. 2016) is a project conceived at the MIT Media Laboratory in 2009. Its development involves collaboration between the Media Lab, the University of Stirling, and Sitekit Solutions Ltd. It is accessible by an API available online, but the exact tool used in the present study is the Python package "senticnet" (ver. 1.0.1).

The package is not just a sentiment lexicon. It offers several useful options. They are available mostly for individual words, but some phrases may also be queried. Among other things, one may obtain "moodtags", such as "#joy" or "#admiration", or the so-called "sentics", which are values for qualities such as "sensitivity", "attention", "aptitude" and "pleasantness". Most relevant to the present study, however, are the attributes "polarity value" and "polarity intensity". The former is a binary sentiment value for a given word (positive or negative), and the latter refers to a similar result on a gradable scale of –1 (extremely negative) to +1 (extremely positive). "Polarity intensity" is, therefore, the feature which has been used for the current purposes.

The application of SenticNet into the analysis described in Section 4 is fairly straightforward (see Figure 1). Each text from the validation materials described in Section 3.3 was pre-processed by removing punctuation and performing word tokenization using the Python "nltk.word_tokenize" module. Next, polarity intensity was obtained for each word with the use of the "senticnet" Python package. If any of the words were not recognized, the default value of 0.0 was recorded. Finally, all key statistical instruments crucial to the current project were computed for each text.

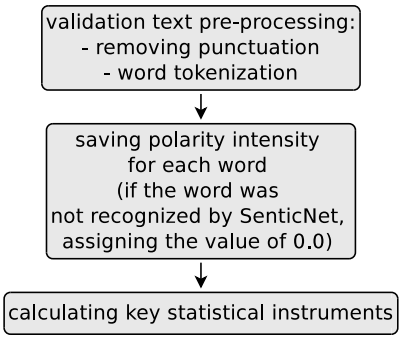

Figure 1. Implementation of SenticNet in the present analysis

3.2.2 SentiStrength Lexicon

SentiStrength is a stand-alone program with a graphic user interface, but other versions of the software are also available. One is an online tool, and another is a "Java version", which is recommended for commercial use and is accessible from the command line. SentiStrength has been described and evaluated in Thelwall (2017), Thelwall et al. (2010, 2012, 2013) and Thelwall and Buckley (2013). It has also been used in numerous research projects.

The use of SentiStrength in the current study involves only the main sentiment lexicon included in the set of resources attached to the program. The lexicon is a tab separated value file with a list of English lexical items and the corresponding sentiment values on a scale of –5 to +5. The only aspect which makes the inclusion of the lexicon in the current analysis less than straightforward is the fact that a large number of the lexical items listed are inflectional or derivational bases rather than final English forms. The items meant to be the bases are marked with an asterisk at the end. For this reason, the adaptation of the lexicon for the purposes of the present study required some additional procedures (see Figure 2). The pre-processing stage of the validation materials was standard and involved removing punctuation and word tokenization using the Python "nltk.word_tokenize" module. What is different from other cases, however, is the division of the lexicon into two separate parts. All the lexical items which were "final forms" were collected in one file, and the items which were inflectional of derivational bases were saved in another file. The corresponding scoring was also saved in these files. For each text from the validation materials described in Section 3.3, all the words were searched in the first file. If any word was found, the corresponding scoring was recorded. Next, a similar search was done in the

second file, but this time the results were collected not only for identical items, but also for cases in which a given word began in the same way as any of the inflectional or derivational bases. For instance, the word "abandoned" would be recognized as a possible form derived from the base "abandon*", present in the second file. Finally, the key statistical instruments defined in Section 3.1 were calculated for each text.

3.2.3 SentiWordNet (SWN)

SentiWordNet is a tool designed to be used in sentiment classification and opinion mining. It has been described in Baccianella et al. (2010), Esuli and Sebastiani (2006, 2007) and Kreutzer and Witte (2013), and applied in numerous research projects. As its name suggests, it was built using WordNet, which is a huge lexical database of English (Fellbaum 1998; Miller 1995).

SentiWordNet may be downloaded directly from "sentiwordnet.isti.cnr.it", but the version used in the present study is the module included in the Python NLTK platform (version 3.2.4) (Bird et al. 2009). Because of the rather complex structure of WordNet, application of SentiWordNet in the current analysis was more complex than the procedures used for other lexicons (see Figure 3). After importing the validation materials described in Section 3.3, punctuation was removed and word tokenization was performed using the Python "nltk.word_tokenize" module. After that, part-of-speech tagging was conducted with the use of "nltk.pos_tag". Next, function words (or "stop words", as they are referred to in NLP) were removed. The reason for this choice is the fact that, even if such items are assigned sentiment values, their interpretation depends almost entirely on context and none of the lexicons in this study takes any pragmatic aspects of texts into account. Then, lemmatization was performed using the "WordNetLemmatizer" class imported from the "nltk.stem.wordnet" module. This operation was necessary, because in the next step the SentiWordNet "senti_synset" method was used, and it recognizes correctly only uninflected forms. The method returns two separate numeric scores. Both are values between 0 and 1. The one called "PosScore" indicates the degree to which a given word is positive, and the one called "NegScore" shows the level of negative associations. Because of this rather uncommon scoring solution, the key statistical instruments defined in Section 3.1 had to be calculated differently than in other cases. Perhaps the best way to describe the exact procedure employed is to give an example. In "this intriguing girl is beautiful, but also mischievous and dangerous" some parts of the expression are positive and others are negative. The results obtained in SWN for this example are presented in Table 2.

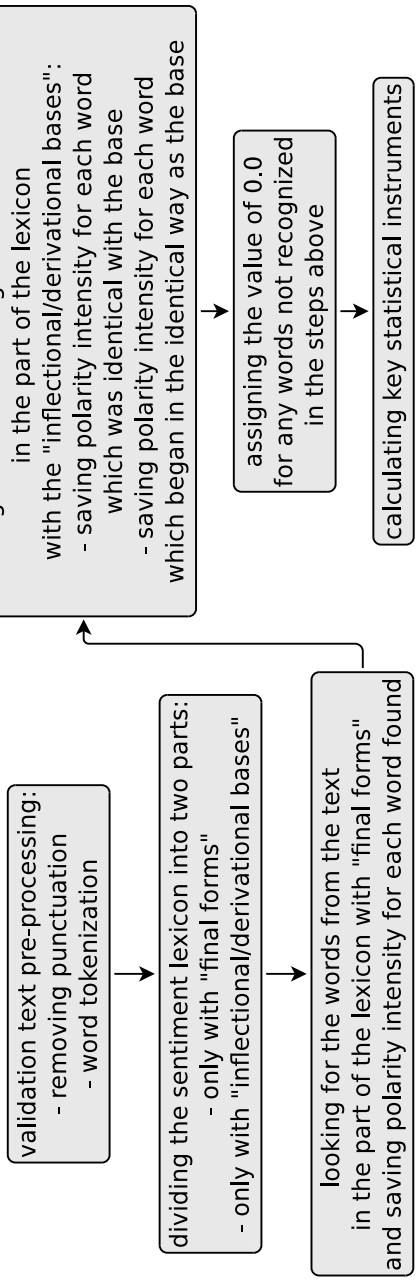

Figure 2. Implementation of SentiStrength Lexicon in the present analysis

Table 2. Results obtained in SWN for an example text containing both positive and negative lexical items

	this	intriguing	girl	is	beautiful	but	also	mischievous	and	unpredict-able
positive scores	0.0	0.5	0.0	0.0	0.75	0.0	0.0	0.0	0.0	0.0
negative scores	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.25	0.0	0.625

SUM would be measured as the difference between positive and negative scores, so (0.5 + 0.75) – (0.25 + 0.625) = 0.375. MEAN1 and MEAN2 could, however, be calculated in at least two different ways:

[1] by obtaining the mean from all positive and negative scores. MEAN1 would be calculated in the following way:
((0.5 + 0.75) – (0.25 + 0.625)) / 4 = 0.09375
MEAN2, which includes "zeroes" for words which have not been found in the lexicon, would be obtained as follows:
((0.5 + 0.75) – (0.25 + 0.625)) / 20 = 0.01875

[2] by obtaining the mean separately for positive scores and separately for negative scores, and then calculating the difference between the two results. MEAN1 would be calculated in the following way:
((0.5 + 0.75) / 2) – ((0.25 + 0.625) / 2) = 0.1875
MEAN2, similarly as before, would involve larger denominators:
((0.5 + 0.75) / 10) – ((0.25 + 0.625) / 10) = 0.0375

After performing correlation tests, it became clear that the first method was more effective. Therefore, MEAN1 and MEAN2 for SWN were calculated in the way described in [1] above.

In computing MEDIAN, only the option involving the whole set of positive and negative scores was considered (again, option [1]). This key statistical instrument requires larger datasets to indicate the middle value in a meaningful way, so there was no sense in splitting the calculations into two parts. In our example, MEDIAN = 0.125. Likewise, LAV was also obtained from the whole set of positive and negative scores. In the example under discussion LAV is 0.75.

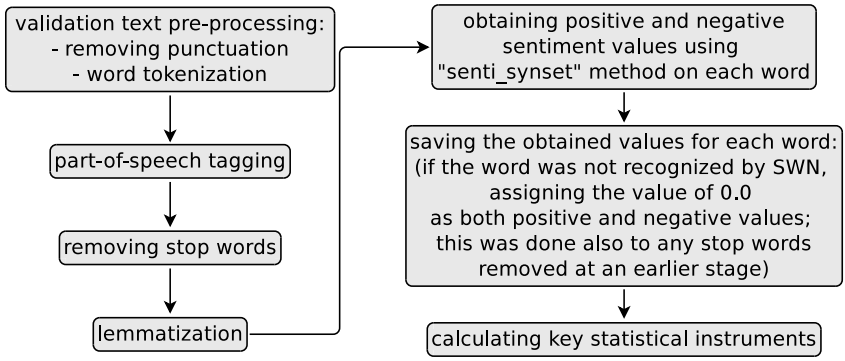

Figure 3. Implementation of SentiWordNet in the present analysis

3.2.4 UMass Amherst Linguistics Sentiment Corpora (UMALSC)

UMass Amherst Linguistics Sentiment Corpora (Constant et al. 2009; Potts – Schwarz 2008) is a collection of n-gram counts extracted from a large number of online product reviews in four languages: Chinese, English, German, and Japanese. The part which is of interest to the present study are the eight datasets for the English language. Four of them contain statistics on bigrams, and the other four focus on unigrams. Although both unigrams and bigrams could be used in this study, for the sake of simplicity only unigrams were utilized. The four files were created on the basis of the following sources: 1) English Amazon book reviews; 2) English Amazon book summaries; 3) English Tripadvisor.com reviews; and 4) English Tripadvisor.com summaries.

Table 3. An example of the structure of UMALSC unigrams counts

Token	Rating	TokenCount	RatingWideCount
absurd	1	35	570687
absurd	2	27	512643
absurd	3	14	767958
absurd	4	20	1513776
absurd	5	48	4769921
abundance	1	8	570687
abundance	2	7	512643
abundance	3	22	767958
abundance	4	43	1513776
abundance	5	109	4769921

An example of the structure of the files is shown in Table 3. For each word type, token counts are provided for 5 ratings. The ratigs are on a gradable scale of 1 (very negative) to 5 (very positive). Additionally, the total token count for each rating is also provided. Such raw data needed to be processed in order to obtain a single sentiment score for each word type. The solution chosen involved two stages. Firstly, the four datasets were concatenated into one. Each word type present in any of the four files was searched for in the other three files. If it was present only in this dataset, the token counts were just copied to the concatenated file. However, if a given word type was found in more than one dataset, its token counts were added and the sum was saved in the concatenated file instead. Secondly, a single, unidimensional measure of sentiment for each word type was calculated using the formula presented below. x represents "token count" for a given rating, and w is the weight assigned to each rating ($w_1 = -2$, $w_2 = -1$, $w_3 = 0$, $w_4 = 1$, $w_5 = 2$).

$$\frac{\sum_{i=1}^{5} x_i \, w_i}{5 \sum_{i=1}^{5} x_i}$$

The resulting sentiment lexicon was implemented in the present analysis in the same way as the Vader Lexicon described below (see Section 3.2.5 and Figure 4).

3.2.5 Vader Lexicon (VL)

The lexicon described in this section is included in the sentiment analysis tool known as "Valence Aware Dictionary and sEntiment Reasoner" or VADER (Hutto – Gilbert 2014). The tool is available as a Python library and it involves both a lexicon and a rule-based sentiment analysis algorithm. Nevertheless, this study focuses only on the former component, which will be referred to as "Vader Lexicon" or VL. The lexicon is a tab delimited file. It provides sentiment ratings on a scale of –4 (very negative) to +4 (very positive) for over 7000 word types created on the basis of ratings provided by multiple independent human judges. The lexicon is especially attuned to social media contexts but may be useful for sentiment analysis in other domains.

The way in which Vader Lexicon has been applied in the present study is summarised in Figure 4. In the validation texts discussed in Section 3.3 all the punctuation was removed and word tokenization was

performed with the use of the "nltk.word_tokenize" module. Next, part-of-speech tagging was conducted using "nltk.pos_tag". Then, function words were removed from the texts. Since VL does not include such lexical items, this step was optional and it was performed solely for increasing the speed of processing. After that, the "WordNetLemmatizer" class imported from the "nltk.stem.wordnet" module was used for lemmatization. Sentiment values were obtained in a three-step procedure. If a given word type was found in the lexicon, the value was assigned to it directly. If the word type was not found, however, the corresponding lemma was searched for in VL. This step maximized the number of lexical items scored and potentially improved the general performance of the lexicon. Finally, if the word type was not found in any of the two previous steps, the default value of 0.0 was assigned to it. The same was done to any function words removed at an earlier stage.

The key statistical instruments for each text were calculated in the standard manner described in Section 3.1.

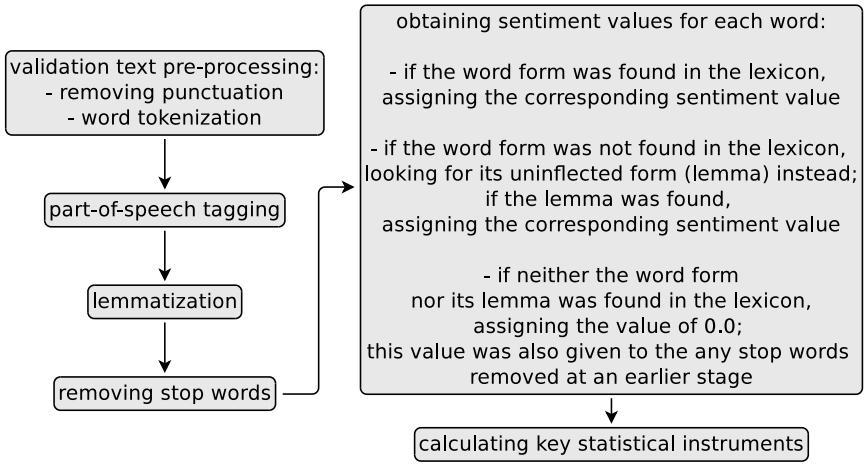

Figure 4. Implementation of Vader Lexicon in the present analysis

3.2.6 Sentiment and Emotion Lexicons created by the National Research Council Canada (NRC)

The tools offered by the National Research Council Canada include a variety of different sentiment and emotion lexicons. Nine of these lexicons involve numerical scoring for the dichotomy "positive" vs. "negative" and are appropriate for current purposes. They are briefly described below.

[1] Sentiment Composition Lexicon for Negators, Modals, and Degree Adverbs (SCL-NMA) has been described in Kiritchenko and Mohammad (2016b). The lexicon comprises 1621 single words and 1586 phrases. The phrases were formed by combining single words with an auxiliary verb, a degree adverb, a negator, or a combination of those. Each single word or multiple-word phrase has been given a sentiment score on a scale of –1 (very negative) to +1 (very positive). The scores were obtained through crowdsourcing.

[2] SemEval-2015 English Twitter Lexicon (ETL) has been discussed in Kiritchenko et al. (2014). It contains 1515 single words and two-word phrases. All of them have been taken from English Twitter. The two-word expressions are composed of words proceeded by negators. Each single word and two-word phrase has been given a sentiment value on a scale of –1 (very negative) to +1 (very positive). As in the previous case, the scores were obtained through crowdsourcing.

[3] Sentiment Composition Lexicon for Opposing Polarity Phrases (SCL-OPP) (Kiritchenko – Mohammad 2016a, 2016c) consists of 1178 unigrams, bigrams and trigrams which were taken from tweets. The two-word and three-word phrases contain at least one positive word and at least one negative word. Again, the sentiment scoring involves a scale of –1 to +1 and it was obtained through crowdsourcing.

[4] NRC Hashtag Sentiment Lexicon (HSL) (Kiritchenko et al. 2014; Mohammad et al. 2013; Zhu et al. 2014) consists of 50836 unigrams and 245920 bigrams. A file with pairs of unigrams and bigrams is also available, but it has not been used in this paper. The unigrams and bigrams were automatically generated from 775000 tweets with sentiment-word hashtags. Each unigram and bigram has been assigned a sentiment value using the algorithm described in Kiritchenko et al. (2014, p. 732). Most of the scores are between –2 and +3, but in extreme cases they reach values above 8.

[5] Hashtag Affirmative Context Sentiment Lexicon and Hashtag Negated Context Sentiment Lexicon (HSL-AFF-NEG) (Kiritchenko et al. 2014; Mohammad et al. 2013; Zhu et al. 2014) contains 43904 unigrams and 174904 bigrams. For some unigrams and bigrams it was indicated that they were taken from negated contexts. For the purposes of the current study, all such cases were removed from the lexicon. The

unigrams and bigrams were generated from the source used in HSL and sentiment scores were assigned using the same method as in the previous case.

[6] Emoticon Lexicon (EL) (Kiritchenko et al. 2014; Mohammad et al. 2013; Zhu et al. 2014) was automatically generated from 1.6 million tweets with emoticons. As in the case of HRC, the lexicon is divided into unigrams (62447 words), bigrams (641737 two-word phrases) and pairs of unigrams and bigrams, but again, the latter file has not been utilized in this study. Sentiment values were assigned using identical methods as in the previous two cases.

[7] Emoticon Affirmative Context Lexicon and Emoticon Negated Context Lexicon (EL-AFF-NEG) (Kiritchenko et al. 2014; Mohammad et al. 2013; Zhu et al. 2014) is based on the same materials as EL. Moreover, the same methods were used in assigning sentiment scores. The lexicon comprises 55054 unigrams and 262142 bigrams. As in the case of HSL-AFF-NEG, some items were marked for having been taken from negative contexts. For the current purposes, such unigrams and bigrams were removed from the lexicon.

[8] Yelp Restaurant Sentiment Lexicon (YRSL) (Kiritchenko et al. 2014) contains 39232 unigrams and 268303 bigrams and was automatically generated from customer reviews from the Yelp Phoenix Academic Dataset available at "http://www.yelp.com/dataset_challenge". Sentiment scores were assigned automatically using the same methods as in the previous four cases. Again, some examples were removed from this lexicon because the context in which they were originally used involved negation.

[9] Amazon Laptop Sentiment Lexicon (ALSL) (Kiritchenko et al. 2014) was generated from reviews on laptops and notebooks collected from "Amazon.com". The lexicon includes 26561 unigrams and 149118 bigrams. Sentiment scores were assigned in the same manner as in the previous five cases. Also, the unigrams and bigrams which were marked for coming from negated contexts were removed as in the case of HSL-AFF-NEG, EL-AFF-NEG and YRSL.

A summary of the processing used in preparing NRC lexicons for the current study is presented in Figure 5. The first two steps have been discussed in

the individual descriptions above. The third stage, however, requires further explanation. After removing punctuation, some expressions were duplicated and a Python script was written to merge them into one. The resulting sentiment scoring was the mean of the scores for all the instances of the duplicated expression. Because of the large size of NRC lexicons, this part of processing was performed at the Mathematical Modelling Laboratory at Jan Kochanowski University in Kielce, Poland.

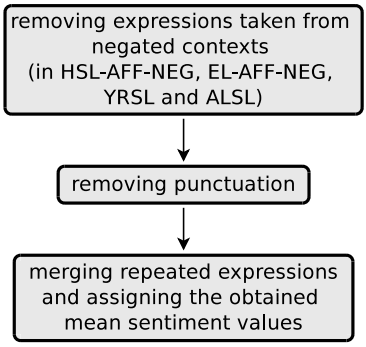

Figure 5. Processing involved in preparing NRC lexicons for the current project

On the basis of the NRC tools described above, a new lexicon was created. It will be referred to as NRC Combined or "NRCC". It is an aggregate of all NRC lexicons. The process of generating it involved copying all the expressions and their sentiment scores into one file. After that, duplicated expressions were merged. The assigned sentiment scores were the means calculated from all the instances of a given, duplicated expression. The task was computationally demanding and, again, the processing was performed at the Mathematical Modelling Laboratory at Jan Kochanowski University in Kielce, Poland.

Figure 6 shows the implementation of all NRC lexicons in the present analysis. After standard preprocessing involving punctuation removal and word tokenization, part-of-speech tagging was performed on each validation text using the Python "nltk.pos_tag" package. Next, lemmatization was conducted with the "WordNetLemmatizer" class imported from the "nltk.stem.wordnet" module. Sentiment values were obtained in a procedure more complex than those of the cases described previously. NRC lexicons prepared for the present analysis contained unigrams, bigrams and trigrams. In each validation text, trigrams were searched first. If any were found, the corresponding sentiment values were recorded and the trigrams were removed from the text. Next, bigrams were searched. If any were found, the

sentiment values were saved and the bigrams were removed from the text. The same procedure was repeated for unigrams. Additionally, the lemmas of the remaining words were searched in the unigrams part of a given dictionary and if any were found, the corresponding sentiment values were recorded. Finally, the value of 0.0 was assigned to any remaining lexical items and the statistical instruments under analysis were calculated in the standard way described in Section 3.1.

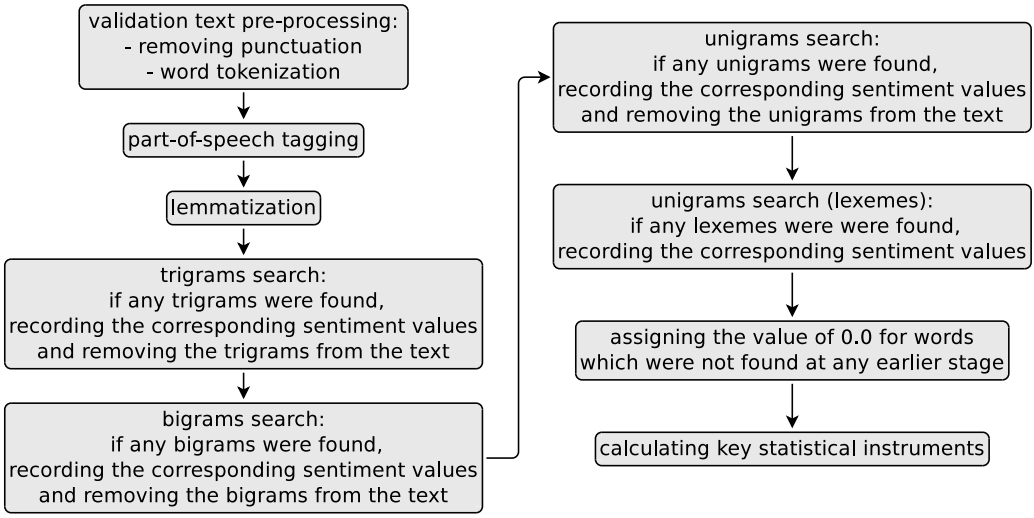

Figure 6. Implementation of NRC lexicons in the present analysis

3.3 Validation materials

The validation materials used in this study were taken from two independent projects on sentiment analysis. Four of the datasets come from the VADER project described in Hutto and Gilbert (2014) (see also Section 3.2.5). The other six were downloaded from the official website of SentiStrength and they are characterized in Thelwall et al. (2012) (see also Section 3.2.2).

A summary of the validation materials is presented in Table 4. It must be stressed, however, that this summary provides statistics on the way the data were used in this study, rather than on their original characteristics. The author could not get access to some parts of the raw materials and some examples were excluded in a few datasets. (Similar problems were encountered in previous studies involving the currently used test data, e.g. Ribeiro et al. 2016).

The materials differ in terms of the number of texts. The dataset with the smallest number of snippets is "BBC forum posts" (693 texts), the one with the largest number is "Rotten Tomatoes movie reviews" (over 10000 texts) and the average number of texts for all 10 datasets is 3349.6. A quick glance at Table 4, however, suggests, that it is also necessary to take into account other aspects of the excepts, such as the average number of words in each fragment in a given dataset[1]. For instance, the mean text length in "Rotten Tomatoes movie reviews" (18.83 words) is much shorter than the average text length in "BBC forum posts" (60.79). Therefore, a statistical instrument which considers the overall number of words of each dataset should be used. Indeed, "Rotten Tomatoes movie reviews" is the largest of the test datasets used in this study. It contains almost 200000 words. The smallest, on the other hand, is "MySpace comments", with 20001 words. The average number of words for all 10 datasets is 66304.3 and the sum of all the words in the data amounts to 663043.

Not only are the validation materials used in this paper extensive, but they represent different types of social Internet environments. The materials obtained from the SentiStrength website concentrate on various comments and posts (Thelwall et al. 2012). "BBC forum posts" involve discussions about various serious topics, "Digg posts" represent news commentaries, "Runners World forum posts" include messages exchanged by a common-interest group, "Twitter posts 2" are public blog broadcasts and "YouTube comments" represent comments on resources available at "youtube.com." The test materials downloaded from Vader's GitHub repository, on the other hand, focus on reviews ("Amazon product reviews", "Rotten Tomatoes movie reviews") and opinion news articles ("New York Times opinion editorials"). "Twitter posts 1" are similar to "Twitter posts 2", but according to the description on Vader's website, they are "tweet-like" texts "inspired" by tweets rather than unaltered messages obtained directly from "twitter.com".

Each text in the 10 datasets was rated for sentiment value by (a) human participant(s). What is most crucial, however, is the fact that the ratings are not polarity-based, but valence-based. Instead of classifying the texts as positive or negative, a gradable scale was used. In the case of the Vader datasets, the scale was from –4 (extremely negative) to +4 (extremely

[1] The number of words in each dataset was calculated in Python using the "nltk.tokenize" package. Emoticons and other symbols which are not part of the Roman alphabet were excluded.

Table 4. Statistics on the validation materials

dataset	abbreviation	source	type of texts	number of texts	mean text length (number of words)	sd of text length (number of words)	number of words
Amazon product reviews	amazon_reviews	Vader	customer reviews	2693	15.96	10.37	42969
Rotten Tomatoes movie reviews	movie_reviews	Vader	movie reviews	10605	18.83	8.69	199726
New York Times opinion editorials	nyt_editorial	Vader	opinion editorials	5181	17.42	8.71	90322
Twitter posts 1	tweets	Vader	tweets	4198	13.41	6.69	56308
BBC forum posts	bbc	SentiStrength	social media comments	693	60.79	72.09	42311
Digg posts	digg	SentiStrength	social media comments	1077	31.45	44.23	33873
MySpace comments	myspace	SentiStrength	social media comments	1041	19.21	25.10	20001
Runners World forum posts	rw	SentiStrength	social media comments	1046	63.61	68.44	66545
Twitter posts 2	twitter	SentiStrength	social media comments	3555	14.94	6.39	53142
YouTube comments	youtube	SentiStrength	social media comments	3407	16.98	17.83	57846
all materials mean				3349.6	27.26	26.854	66304.3
all materials sum				33496			663043

positive). Any values in between represented more moderate attitudes, with 0 indicating neutrality. Similarly, SentiStrength materials were graded on a scale of 1 to 5, but separately for positive and negative emotions. For instance, a text regarded as extremely positive would be given 5 on the "positive emotion scale" and 1 on the "negative emotion scale". In the present study, "negative scores" were deducted from "positive scores" and the resulting value was assumed to represent the sentiment value of a given text. For example, if the score on the "positive emotion scale" was 5 and on the "negative emotion scale" was 1, the score used for the present paper was 4. Consequently, the range of the scoring was from –4 to +4, just as in the previous case.

3.4 Statistical tests

The analysis described in Section 4 involves performing tests for association between paired samples, using Pearson's product moment correlation coefficient. The independent variables are datasets with the key statistical instruments obtained for each text in the validation materials according to each lexicon. For instance, for the entire "Amazon product reviews" collection there are as many as 75 such datasets (5 types of key statistical instruments × 15 lexicons). In each case, the dependent variable is the corresponding human scoring. The correlation coefficients obtained are further tested with the use of ANOVA and Tukey HSD post-hoc analysis. The choice of these parametric methods is based on the observation that both the normality condition and the equal variance condition are not severely violated (see Figure 7).

It is worth noting that the independent variables involve numeric results on different scales. This stems from the fact that the sentiment lexicons themselves use different scoring ranges. Nevertheless, no attempt has been made to normalize the data since it is not really a problem for the correlation tests as long as the scale is the same for all the data in a given set. The resulting correlation coefficients will be the same, no matter what the range of the scoring scale is.

All the statistical tests were performed using R (R Development Core Team 2013).

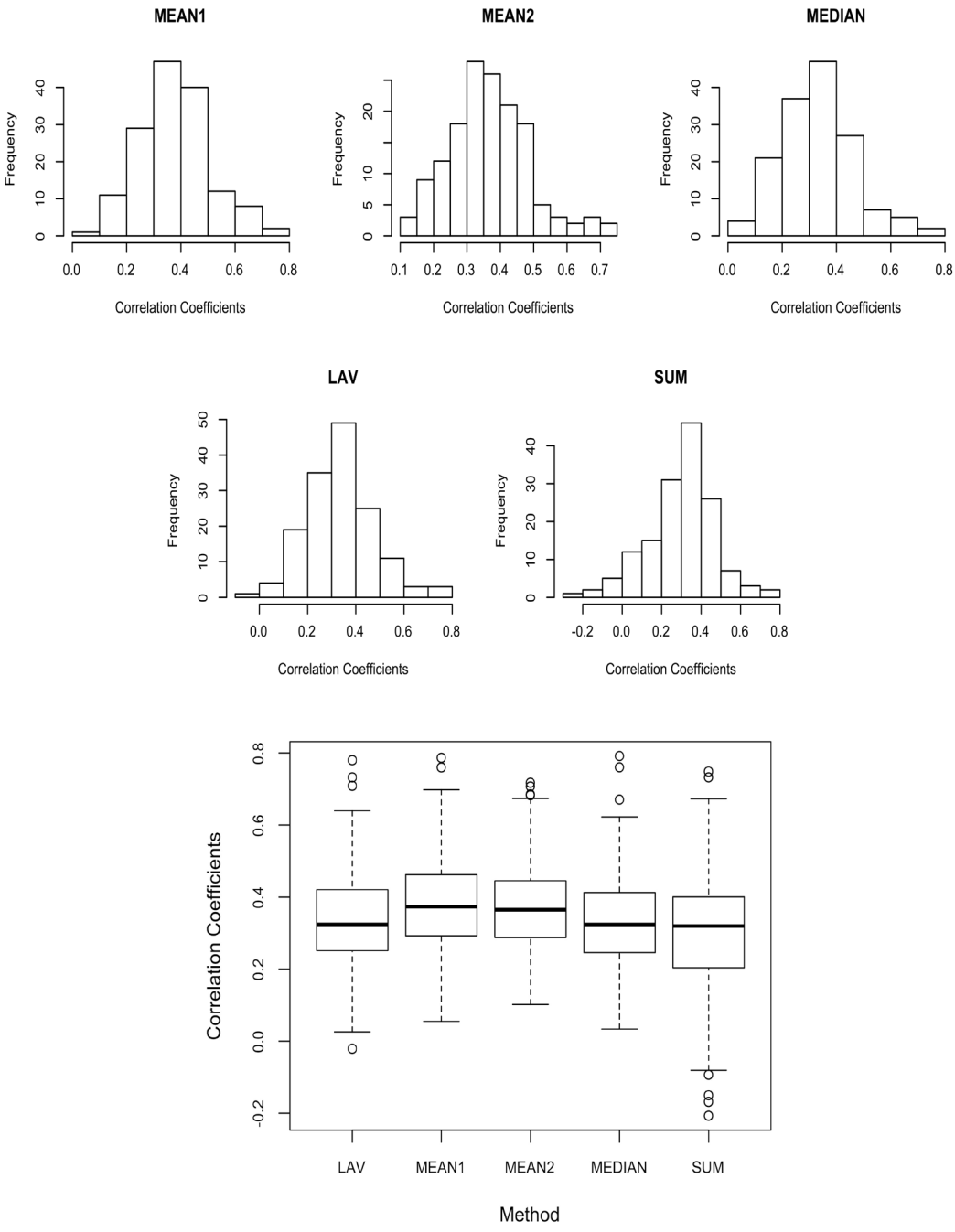

Figure 7. Histogram and boxplot of the correlation coefficients obtained for data including negation (similar patters were observed for data excluding negation)

4. Results

The results are summarised separately for tests performed on entire validation datasets (Section 4.1) and tests conducted on the validation materials divided into smaller samples (Section 4.2).

4.1 Results based on entire validation sets

The general results of the analysis performed in this study are presented in Table 5. They are based on 750 correlation tests (5 types of key statistical instruments × 15 lexicons × 10 validation sets). The values in the column "average correlation for all texts" were obtained by averaging correlation coefficients for all validation sets with all the texts they include. It is immediately visible that the method with the highest mean correlation is MEAN1 (0.38189). The second most successful measure is MEAN2, with the result of 0.36916, which is 0.01273 less than in the case of MEAN1. MEDIAN and LAV performed on a similar level. The mean correlation for both is around 0.33. Finally, the method with the lowest result is SUM. Its mean correlation is only 0.29631, which is almost 0.1 less than the mean correlation obtained for MEAN1.

Table 5. Average correlation for key statistical instruments based on entire validation sets

method	average correlation for all texts	average correlation excluding texts with negation
MEAN1	0.38189	0.40189
MEAN2	0.36916	0.38673
MEDIAN	0.33607	0.35716
LAV	0.33121	0.35905
SUM	0.29631	0.34213

An ANOVA for the data discussed above was performed. It revealed a statistically significant difference between at least two groups ($F(4,745) = 8.654$, $p < 0.0001$), so a Tukey HSD post-hoc analysis was also conducted. The pairs whose comparison yielded p-values below the alpha level of 0.05 are listed below:

- MEAN1 – MEDIAN ($p = 0.0392$)
- MEAN1 – LAV ($p = 0.016$)
- MEAN1 – SUM ($p < 0.0001$)
- MEAN2 – SUM ($p < 0.0001$)

These results clearly indicate that MEAN1 is, in fact, the most effective method of those under investigation. The only possible exception is MEAN2, whose lower ranking has not been statistically substantiated. Besides the four pairs above, no other comparison indicated that the differences were statistically significant. One of the confounding factors which might explain this is the fact that some of the texts in the validation tests involve negation, which is a broadly discussed issue in sentiment analysis (see the last paragraph in Section 1). Many tools, such as Vader or Sentistrength, are designed to cope with it, but the simplistic processing used in calculating the five key statistical instruments in this study does not deal with this problem at all. Therefore, a more appropriate solution is to use validation materials without texts involving negation.

A script was written in Python and all examples with negation were removed. Next, another series of correlation tests was conducted. The results obtained are summarised in Table 5 in the column "average correlation excluding texts with negation". They are based on exactly the same number of tests (5 types of key statistical instruments × 15 lexicons × 10 validation sets = 750 correlation tests), but this time each validation set is shorter and does not include any negated sentences. The average correlation coefficients obtained are higher by approximately 0.02, with the exception of the result for SUM, which is higher by almost 0.05. The relative ranking of the methods, however, does not change in any significant way. Again, the statistical instrument which produces the best results is MEAN1, with MEAN2 close behind, followed by MEDIAN and LAV. The result for SUM is still the worst, although the gap between it and the other methods is smaller than in the previous analysis. An ANOVA performed on these data indicated statistically significant difference between at least two groups ($F(4,745) = 4.494$, $p = 0.0014$), so a Tukey HSD post-hoc analysis was conducted once again. The comparison of each possible pair yielded results very similar to those of the previously performed Tukey HSD. In fact, the p-values obtained are slightly higher. The ones still indicating statistical significance are listed below:

- MEAN1 – MEDIAN (p = 0.0.0463)
- MEAN1 – SUM (p = 0.0022)
- MEAN2 – SUM (p = 0.0473)

The result for the pair MEAN1 – LAV is marginally significant (p = 0.0634). All the rest of the comparisons indicate that the differences cannot be statistically confirmed.

4.2 Results based on validation sets divided into smaller samples

A factor which is important in the tests performed thus far is the size of samples. In the previous calculations, the mean correlation coefficient for each method has been computed on the basis of 150 correlation tests (15 lexicons × 10 validation sets). The ANOVA and Tukey HSD performed later did not take into account the actual size of the samples on which the correlation tests were performed. It is, however, possible to divide the 10 validation datasets into smaller sets. In this way the number of the actual correlation tests could be increased substantially.

Table 6. Division of validation materials into smaller samples

dataset	including negation		excluding negation	
	number of texts	number of samples	number of texts	number of samples
Amazon product reviews	2693	27	2044	21
Rotten Tomatoes movie reviews	10605	106	8071	81
New York Times opinion editorials	5181	52	4310	43
Twitter posts 1	4198	42	3056	31
BBC forum posts	693	7	375	4
Digg posts	1077	11	704	7
MySpace comments	1041	11	881	9
Runners World forum posts	1046	11	655	7
Twitter posts 2	3555	36	3056	31
YouTube comments	3407	34	2844	29
sum	33496	337	25996	263

Statistics coursebooks (e.g. Rumsey 2003) usually suggest that the minimum sample size for obtaining reliable results is around 30. Since the validation datasets used in this study are large, samples of 100 were eventually chosen, if the final group of texts totalled at least 30, it was included in the analysis. The way in which each validation dataset was divided into smaller sets is presented in Table 6. For instance, "Amazon product reviews" contains 2693 texts. Consequently, 26 samples of 100 texts were obtained plus one

final sample with 93 texts. Since this final sample is larger than the minimum of 30, it has been included in the analysis, so the ultimate number of samples for this validation dataset is 27. However, in the dataset "Rotten Tomatoes movie reviews", the last sample was ignored because it contains only 5 texts.

After dividing the validation materials into smaller samples, as many as 25275 correlation tests were performed (5 types of key statistical instruments × 15 lexicons × 337 validation sets). This time, each mean correlation coefficient for each key statistical instrument was calculated on the basis of 5055 measurements (15 lexicons × 337 validation sets). The results obtained are summarised in Table 7 in the column "average correlation for all texts". The ranking of the methods is identical to the hierarchy observed before. MEAN1 is the most efficient statistical instrument, closely followed by MEAN2. On this occasion, however, the difference is smaller than before (only about 0.006). The results for the other three methods are very similar to each other, but clearly lower than the average correlations obtained for MEAN1 and MEAN2. An ANOVA performed on these data revealed a statistically significant difference between at least two groups ($F(4,25270) = 106.4$, $p < 0.0001$), so a Tukey HSD post-hoc analysis was conducted. Here, the majority of the differences between the results are statistically significant, with the exception of the four pairs listed below.

Table 7. Average correlation for key statistical instruments based on validation sets divided into smaller samples

method	average correlation for all texts	average correlation excluding texts with negation
MEAN1	0.35900	0.38138
MEAN2	0.35296	0.37338
MEDIAN	0.31448	0.33712
LAV	0.30926	0.33638
SUM	0.30954	0.33916

- MEAN1 – MEAN2 (p = 0.4713)
- MEDIAN – LAV (p = 0.6338)
- MEDIAN – SUM (p = 0.6877)
- LAV – SUM (p = 0.9999)

These analyses show that both MEAN1 and MEAN2 are more effective statistical instruments than MEDIAN, LAV and SUM. No other differences, however, have been confirmed. The same conclusions can be drawn from

the analysis excluding texts involving negation. Although the average correlation coefficients summarised in Table 7 are higher by over 0.02, the ranking of the methods, as well as the relative differences between the way they performed, is identical to the analysis with "all texts". Indeed, a Tukey HSD post-hoc analysis has revealed that only differences between the same four pairs (MEAN1 – MEAN2, MEDIAN – LAV, MEDIAN – SUM, LAV – SUM) cannot be confirmed statistically.

5. Conclusion

Lexicon-based tools for sentiment analysis frequently involve complex rule-based sentiment analysis algorithms. These algorithms aim at compensating for various linguistic phenomena, such as negation and intensification. Ultimately, they summarise the analysis performed by providing either a specific category (e.g. "positive" or "negative") or a numeric polarity score.

In the present study, an investigation was made into key statistical instruments that may be used to obtain such final results. The data analysed indicate that, other things being equal, the mean is more effective than the median, the largest absolute value, or the sum in obtaining numeric polarity scores similar to the scores provided by human participants. This conclusion may be useful in improving software packages performing lexicon-based sentiment analysis. If there is no compelling reason to use other statistical instruments in the calculation of the final score, the mean is the best option. Such a decision may also be made by the end users of tools which offer a variety of options for calculating final sentiment scores (e.g. SentiStrength).

As far as the exact way in which the mean should be calculated, no definitive answer can be offered. Two different methods were tested, one excluding lexical items not found in a given sentiment lexicon and the other including such items. Neither of the two methods was clearly more efficient than the other.

REFERENCES

Abbasi, A. – A. Hassan – M. Dhar
2014 "Benchmarking Twitter sentiment analysis tools." In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, 26-31.

Aue, A. – M. Gamon
2005 "Customizing sentiment classifiers to new domains: A case study". In: *Proceedings of Recent Advances in Natural Language Processing (RANLP)*. Borovets, Bulgaria, September 17-19, 2005.
Baccianella, S. – A. Esuli – F. Sebastiani
2010 "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." In: *Proceedings of the 7th International Conference on Language Resources and Evaluation* (*LREC),* 2200-2204.
Bird, S. – E. Klein – E. Loper
2009 *Natural Language Processing with Python*. Beijing: O'Reilly.
Borth, D. et al.
2013 "Large-scale visual sentiment ontology and detectors using adjective noun pairs". In: A. Jaimes (ed.) *Proceedings of the 21st ACM International Conference on Multimedia*. New York: ACM, 223-232.
Cambria, E. et al.
2016 "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives". In: *Proceeding of COLING 2016, The 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan, December 11-17 2016, 2666-2677.
Choi, Y. – C. Cardie
2008 "Learning with compositional semantics as structural inference for subsentential sentiment analysis". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, October 2008,* 793-801.
Constant, N. et al.
2009 "The pragmatics of expressive content: Evidence from large corpora", *Sprache und Datenverarbeitung 33* (1-2), 5-21.
Diniz, J.P. et al.
2016 "iFeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis". In: *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*. Association for the Advancement of Artificial Intelligence.
Esuli, A. – F. Sebastiani
2006 "SENTIWORDNET: A publicly available lexical resource for opinion mining". In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. May 2006, Genoa.
2007 "SENTIWORDNET: A high-coverage lexical resource for opinion mining". *Technical Report 2007-TR-0*2. Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche. Pisa, Italy.
Fellbaum, C.
1998 *WordNet: An Electronic Lexical Database*. Cambridge, Mass.; London: MIT Press.

Gonçalves, P. et al.
2013 "Comparing and combining sentiment analysis methods". In: *Proceedings of the First ACM Conference on Online Social Networks*, 27-38.
Hutto, C.J. – E. Gilbert
2014 "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of The Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14)*, 216-255.
Kennedy, A. – D. Inkpen
2006 "Sentiment classification of movie reviews using contextual valence shifters", *Computational intelligence 22* (2), 110-125.
Kiritchenko, S. – S. Mohammad
2016a "Happy accident: A sentiment composition lexicon for opposing polarity phrases". In: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*. Portorož, Slovenia.
2016b "The effect of negators, modals, and degree adverbs on sentiment composition". In: *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*. San Diego, California.
2016c "Sentiment composition of words with opposing polarities". In: *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).* San Diego, California, 1102-1108.
Kiritchenko, S. – X. Zhu – S.M. Mohammad
2014 "Sentiment analysis of short informal texts", *Journal of Artificial Intelligence Research 50*, 723-762.
Kreutzer, J. – N. Witte
2013 *Opinion mining using SentiWordNet*. Semantic Analysis, HT 2013/14. Uppsala University. http://santini.se/teaching/sais/Ass1_Essays_FinalVersion/Kreutzer_Julia_AND_Witte_Neele_SentiWordNet_Neele+Julia_finalversion.pdf, accessed December 2021
Liu, B.
2012 "Sentiment analysis and opinion mining", *Synthesis Lectures on Human Language Technologies 5* (1), 1-167.
Liu, B. – L. Zhang
2012 "A survey of opinion mining and sentiment analysis". In: C.C. Aggarwal – C. Zhai (eds.) *Mining Text Data.* Springer, 415-463.
Miller, G. A.
1995 "WordNet: A lexical database for English", *Communications of the ACM* 38 (11), 39-41.
Mohammad, S.M. – S. Kiritchenko – X. Zhu
2013 "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets". In: *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, June 2013, Atlanta, USA.

Mohammad, S.M. – P.D. Turney
2010 "Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon". In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles, California: Association for Computational Linguistics, 26-34.
2013 "Crowdsourcing a word–emotion association lexicon", *Computational Intelligence* 29 (3), 436-465.
Nielsen, F.Å.
2011 "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs". In: *Proceedings of the ESWC2011 Workshop on "Making Sense of Microposts": Big things come in small packages (2011),* 93-98.
Pang, B. – L. Lee
2008 "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval 2* (1-2), 1-135.
Potts, C. – F. Schwarz
2008 "Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora", *Ms., UMass Amherst*, 1-29.
R Development Core Team
2013 *R: A Language and Environment for Statistical Computing*. [computer software]. Version 3.0.3.
Ribeiro, F.N. et al.
2016 "SentiBench – A benchmark comparison of state-of-the-practice sentiment analysis methods", *EPJ Data Science 5* (23), 1-29.
Rumsey, D.J.
2003 *Statistics for Dummies*. Hoboken, N.J.: Wiley Publishing.
Taboada, M. et al.
2011 "Lexicon-based methods for sentiment analysis", *Computational Linguistics 37* (2), 267-307.
Thelwall, M.
2017 "Heart and soul: Sentiment strength detection in the social web with SentiStrength". In: J. Holyst (ed.) *Cyberemotions: Collective Emotions in Cyberspace.* Berlin: Springer, 119-134.
Thelwall, M. – K. Buckley
2013 "Topic-based sentiment analysis for the social web: The role of mood and issue-related words", *Journal of the Association for Information Science and Technology 64* (8), 1608-1617.
Thelwall, M. – K. Buckley – G. Paltoglou
2012 "Sentiment strength detection for the social web", *Journal of the Association for Information Science and Technology* 63 (1), 163-173.
Thelwall, M. et al.
2010 "Sentiment strength detection in short informal text", *Journal of the American Society for Information Science and Technology 61* (12), 2544-2558.

Thelwall, M. et al.
2013 "Damping sentiment analysis in online communication: Discussions, monologs and dialogs". In: A. Gelbukh (ed.) *Computational Linguisticsand Intelligent Text Processing, 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II*. Springer, 1-12.
Zhu, X. – S. Kiritchenko – S. Mohammad
2014 "NRC-Canada-2014: Recent improvements in the sentiment analysis of tweets". In: P. Nakov – T. Zesch (eds.) *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, 443-447.

Address: Łukasz Stolarski, Uniwersytet Jana Kochanowskiego w Kielcach, Instytut Literaturoznawstwa i Językoznawstwa, ul. Uniwersytecka 17, 25-406 Kielce, Poland.
ORCID code: https://orcid.org/0000-0002-2668-5509