Honkela, Timo; Korhonen, Jaakko; Lagus, Krista; Saarinen, Esa

Five-Dimensional Sentiment Analysis of Corpora, Documents and Words

Published in:
Advances in Self-Organizing Maps and Learning Vector Quantization - Proceedings of the 10th International Workshop, WSOM 2014

DOI:
10.1007/978-3-319-07695-9_20

Published: 01/01/2014

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Honkela, T., Korhonen, J., Lagus, K., & Saarinen, E. (2014). Five-Dimensional Sentiment Analysis of Corpora, Documents and Words. In *Advances in Self-Organizing Maps and Learning Vector Quantization - Proceedings of the 10th International Workshop, WSOM 2014* (pp. 209-218). (Advances in Intelligent Systems and Computing; Vol. 295). Springer. https://doi.org/10.1007/978-3-319-07695-9_20

Five-dimensional sentiment analysis of corpora, documents and words

Timo Honkela[1,2], Jaakko Korhonen[3], Krista Lagus[2,4], and Esa Saarinen[3]

[1]University of Helsinki, Department of Modern Languages, Helsinki
[2]Aalto University School of Science
Department of Information and Computer Science, Espoo
[3]Aalto University School of Science
Department of Industrial Engineering and Management, Espoo
[4]National Consumer Research Centre, Helsinki
Finland
<first>.<second>@aalto.fi

Abstract. Sentiment analysis has become a widely used approach to assess the emotional content of written documents such as customer feedback. In positive psychology research, the typical one-dimensional analysis framework has been extended to include five dimensions. This five-dimensional model, PERMA, enables a fine-grained analysis of written texts. We propose an approach in which this model, statistical analysis and the self-organizing map are used. We analyze corpora from various genres. A hybrid methodology that uses the self-organizing maps algorithm and human judgment is suggested for expanding the PERMA lexicon. This vocabulary expansion can be useful for English but it is potentially even more crucial in the case of other languages for which the lexicon is not readily available. The challenges and solutions related to the text mining of texts written in a morphologically complex language such as Finnish are also considered.

Keywords: Text mining, natural language processing, self-organizing map, independent component analysis, positive psychology, education, life-philosophical lecturing

1 Introduction

Computer-based quantitative methods are becoming more and more popular in the study of complex phenomena in social sciences and humanities. This trend has been strengthened by the fact that many modern analysis methods and tools enable non-reductionistic approaches. Quantitative methods may be useful in qualitative analysis if thousands or even larger number of variables are dealt with simultaneously. This idea is reflected in the representation and analysis of texts as large matrices or tensors. Moreover, computational methods enable modeling and simulation that takes into account the systems nature of real world phenomena [3]. Related research areas include systems intelligence [4] and complexity science [1].

1.1 Background motivation

In this paper, we take first steps in approaching one highly complex phenomenon related to psychology, education and philosophy, namely what kind of changes begin to happen in the students that have attended a course built on life-philosophical lecturing. By life-philosophical lecturing we refer to a particular kind of oral pedagogical practice that uses the lecture situation for the benefit of providing for the listeners an enhanced possibility of life-philosophical reflection [17, 18]. The dominant lecturing practices seek to function as a channel for predetermined knowledge, theories or learning. Then the goal is to make the listeners to adopt the insights, scholarship or philosophy of the lecturer. In contrast, in life-philosophical lecturing "the paramount aim is to facilitate, stimulate and vitalize the participants own life-philosophical thinking in the first-person – his or her use of the reflective mind" [17]. Life-philosophical lecturing is a form of positive philosophical practice and seeks key inspiration from the breakthroughs of the positive psychology movement [22, 21]. Our aim is to be eventually able to measure from texts written by students the changes such a lecturing practice stimulates in them. This article describes our first experiments on the matter.

1.2 Sentiment analysis and the PERMA model

Sentiment analysis of written documents aims to determine the overall polarity of each document of the attitude of the author(s) regarding some topic. Sentiment analysis has become commonplace and it is widely applied, e.g., in business intelligence and in analyzing social media contents [25, 15, 14, 5]. The sentiment of a document is typically calculated as a synthesis of the sentiments of the words and phrases in the document. A straightforward approach is to manually associate a positive or negative value for those words that indicate sentiment. Turney automated this process by calculating the sentiment of a given phrase by comparing its similarity to a positive reference word ("excellent") with its similarity to a negative reference word ("poor") [25].

A typical approach in sentiment analysis is to estimate the polarity of the documents. This one-dimensional measure can be replaced by analyzing multiple factors simultaneously. A straightforward extension is to measure both valence (positive vs. negative) and arousal (activation vs. deactivation). A more refined category system of emotions could include level of interest, enjoyment, surprise, contempt, anger, fear, distress and shame.

In the context of positive psychology research, Seligman has developed the PERMA model that addresses different aspects of wellbeing [21]. The PERMA model includes five components related to subjective well-being: Positive emotion (P), Engagement (E), Relationships (R), Meaning (M) and Achievement (A) [21]. Researchers have gathered a PERMA lexicon that is a collection of words that are associated with each of the components in a positive or negative manner [19].

We propose a way to apply the PERMA model to the analysis of document collections. Furthermore, we suggest a way for complementing the PERMA vo-

cabulary that can be useful especially for other languages than English. The PERMA analysis of texts can be considered at three main levels:

1. PERMA profiling of document collections. This can provide an overall understanding of the nature of different corpora. We analyze the five-dimensional profile of corpora in six different genres.
2. PERMA profiling of individual documents. The second level of analysis is seen to be useful for the lecturer who is provided tools for familiarizing himself with certain aspects of hundreds of long essays written by the students. A related idea has been presented in the context of MOOCs (massive online open courses) [8] for mining student contributions.
3. Comparison of PERMA and non-PERMA words. This analysis can be conducted, for example, in order to find new PERMA word candidates. In this paper, we use the self-organizing map [11] for this purpose.

One way to look at sentiment analysis is to ask first, which are the sentiments that need to be detected, and second, which features in the text reflect said sentiments. While PERMA model provides a theory-driven proposal for a set of such sentiments, as well as seed lists of features, challenges remain. For example, many texts might not have many PERMA features at all. Moreover, translating the PERMA vocabulary to another language leads to additional challenges, since each language might have quite different ways to express for example positivity, and literal translation of words may not be a sufficient method for capturing these.

1.3 Why unsupervised methodology

When using learning methods, information regarding properties of interest (features) or decisions of interest (e.g. class labels) need to be provided to the learning system. Feature selection is generally considered a weak form of importing supervision to a learning system, whereas applying labeled data would constitute a strong form of supervision. One could approach this as a classification problem, by providing a number of manually classified samples to the system.

Instead, in this case we apply a theory-driven perspective for selecting the features, namely the PERMA vocabulary, and then apply an unsupervised learning method, namely the self-organizing map for exploring the outcome. By providing prior knowledge in the feature selection stage the researcher is able to give the learning system information regarding the properties of interest, without having to determine exactly what the outcome should be regarding any specific case, such as a document or a collection. The fact that PERMA vocabulary has been collected already by researchers allows the ready use of unsupervised clustering and visualization methods for any new corpora as well. This suits well in a text mining scenario, where the interest is in finding new, surprising phenomena in the direction of interest of the researcher.

1.4 Preprocessing morphologically highly complex languages

From a morphological point of view, English is a rather simple language. This means that methods that are based on lists of keywords (e.g. [24]) can rely on the idea that there are only a small number of different surface forms of the same basic form or stem. On the other hand, many other languages have more complex morphology. For instance, in Finnish every noun has about 2,000 different inflections and every verb more than 10,000. In addition to this, compounding is very commonplace. Common multi-word phrases in English are often translated as compounds in Finnish. The outcome of the complex morphology is that Finnish has billions of surface word forms which cannot be simply categorized and listed. Fig. 1 shows the seven (out of 99) most common forms of 'merkitys' (meaning) in our essay corpus. In order to deal with the problem of varying word

Word	Translation	Freq.
merkitystä	meaning (as a partial object)	142
merkitys	meaning	101
merkityksen	of the meaning	65
merkityksellistä	of the meaningful	41
merkityksiä	meanings (as a partial object)	36
merkityksellisyyden	of the meaningfulness	34
merkityksellisiä	meaningful (plural, as a partial object)	32
...	...	..

Fig. 1: Examples of different forms for the word 'merkitys' (meaning) in Finnish with the frequency count in our essay corpus.

forms, we have relied on the methodology originally developed by Koskenniemi [12]. We used an open-source implementation of the model [13] to process our corpus. The Omorfi tool transformed each inflected word into its basic form. Due to ambiguities and differences in subtle meanings, the process does not preserve all information when language borders are crossed but details cannot be dealt with here.

2 PERMA profiles of different genres

One can compute a PERMA profile for a document by counting the frequencies of the PERMA words in each component. Our hypothesis was that the PERMA profiles would be different for text corpora that represent different genres. We chose the following kinds of corpora: news feeds from Reuters and Finnish news agency STT, Wikipedia articles on topics that start with the letter A, everyday conversations collected at UC Santa Barbara [2], proceedings of European parliament from 1996 [10], English translation of the fairy tales by Grimm brothers, corporate e-mails messages sent in Enron. The Enron corpus was divided into

three parts to check whether the inter-corpus variation is smaller than intra-corpus variation. The result of the PERMA profile analysis is shown in Fig. 2.

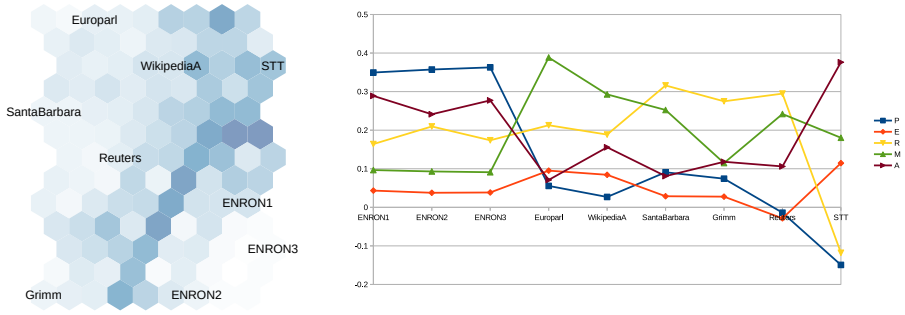

Fig. 2: On the right, relative PERMA profiles for different corpora. Enron e-mails show high values on Positivity and Achievement but low on Meaning, whereas Europarl and Wikipedia are low on Positivity but high on Meaning. On the left, the same PERMA profile information is used to project the corpora on a SOM.

The results indicate that the PERMA vocabulary is able to identify differences among the document collections in a meaningful and informative way. Firstly, the profiles are markedly different for the various corpora. Secondly, the results clearly make sense.

For instance, the news corpora are markedly negative in their content. On the other hand, the Reuters news corpus also scores high on Relationships whereas the Finnish STT scores very high on Achievement. The latter seems to be due to the large proportion of sports news within the STT corpus.

The European parliament corpus obtains high scores on the Meaning dimension and low on Achievement, which may raise a question regarding whether the parliament is concerned enough about achieving any concrete goals.

In a striking contrast are the Enron discussions, which show low Meaning but high Achievement. The emphasis on achievement of concrete goals can be considered natural in a competitive corporate context. However, one is left to wonder whether the low proportion of meaningfulness might have been indicative of upcoming problems, but this is left here as a question for future exploration.

The PERMA analysis over language borders requires further attention. It is probable that phenomena like linguistic polysemy and cultural contextuality influence the results in such as way that fine-tuning of the methodology is necessary. For instance, the four most common positive PERMA words in Finnish in the STT news articles were “voittaa” (to win), “voitto” (victory), “edustaa” (to represent), and “onnistua” (to succeed). This example reminds that the nature of the corpora needs to be carefully concerned.

3 Self-organizing map of sentiment words

In the second case study, we explored the possibility of extending the PERMA vocabulary. This goal was motivated further by the observation that translating the vocabulary to another language necessarily introduces errors due to ambiguity, and is likely to result in an incomplete feature set for that particular language. Thus, research on automatic or semi-automatic means of complementing the feature set is important.

The SOM-based process that we suggest is outlined as a diagram in Fig. 3. The sentiment word list by Hu and Bing [7] was used as an external vocabulary. In the experiment, we used the WikipediaA corpus for calculating the context statistics. We formed word-word context matrices so that each element indicates how many times a sentiment word has appeared in the WikipediaA corpus in the vicinity of a context word. The context words were chosen to first exclude the 100 most common words and then to include the next 2000 words in the order of frequency.

The context window was chosen to be seven words to each direction. This can be characterized as an intermediate choice. Very short context windows emphasize syntactic aspects of the words and document-word matrices work relatively best when the documents are different enough from each other. The resulting matrix was analyzed using the self-organizing map algorithm, presented next. The Self-Organizing Map (SOM) has been used to create word clusters

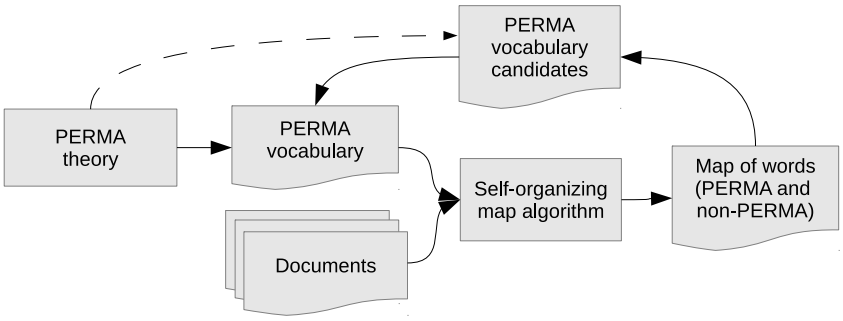

Fig. 3: The process of using the SOM in extending the coverage of a theory-based vocabulary.

automatically from statistical features obtained from corpora [16, 6]. The SOM algorithm produces a topological ordering by mapping the input space to an array of nodes. Each node of a SOM consists of a prototype vector m_i of the same dimension as the input vectors x_i. The nodes are typically organized in the form of a lattice. In an organized map, each input is associated with a prototype in a specific location. The basic idea is that if two input are similar they tend to be close to each other on the map. The SOM is rather similar to clustering

algorithms but does not produce explicit clusters. It rather creates a diagram that aims to "mirror" the high-dimensional data (usually) in two dimensions as faithfully as possible.

In this case, as the data consists of statistical information on the contextual use of words, the end result is map where similar words are close to each other. In the result, several cases may be discussed qualitatively. The positive achievement words (marked by "A+") have been divided into two clusters in the upper and lower left side of the map. Some nearby words such as "progressive" and "renowned" could clearly be considered candidates for extending the A+ lexicon. The number of Relationship (R) and Engagement (E) words in the Hu and Bing word list [7] appeared to be low. The Meaning (M) words form a clear single cluster with the exception of the word "patriotic". These kinds of findings can potentially be used to re-evaluate the PERMA lexicon.

4 Conclusions and discussion

We have applied a five-dimensional PERMA framework and associated vocabulary on performing sentiment analysis on text collections in two different ways.

In the first case, several text collections from different genres were analyzed and their differences observed. We were able to show that the PERMA profiles of the corpora fit with the intuitions related to the types of genres. In addition, the PERMA analysis of the corpora seemed to raise interesting questions such as did Enron fail because it did not concern itself with meaning, or does EU parliament concern itself relatively too much on overall meaning and too little on the achievement of concrete goals to be successful. Based on this it seems that the PERMA framework is promising on the level of corpora and able to highlight interesting differences in the respective discourses.

Our initial objective was to understand more closely the processes that the students go through during life-philosophical lecturing. Due to the challenges related to translating from one language to another, as well as the high number of different word forms in Finnish we found that it would be advantageous to attempt to complement the initial PERMA vocabulary by additional words. We then explored the possibility of doing so using the Word Category Map methodology, where lexical relations of words were used for ordering both PERMA vocabulary and a set of additional words on a two-dimensional display. The ordering is able to identify new candidates for consideration to be added as PERMA features. A fully automatic supervised learning approach could also be used but, on the other hand, the map provides a valuable view on the relational structure of the conceptual space.

Once the feature set is rich enough we expect to be able to extend the PERMA analysis to the full analysis of the PERMA profiles of individual student essays. Due to the complex nature of the philosophical and psychological contents and cognitive and social processes, reductionistic research methods are not easily applicable. It seems, however, that text mining and visualization methods *can* support traditional qualitative analysis [9]. Development of such tools is

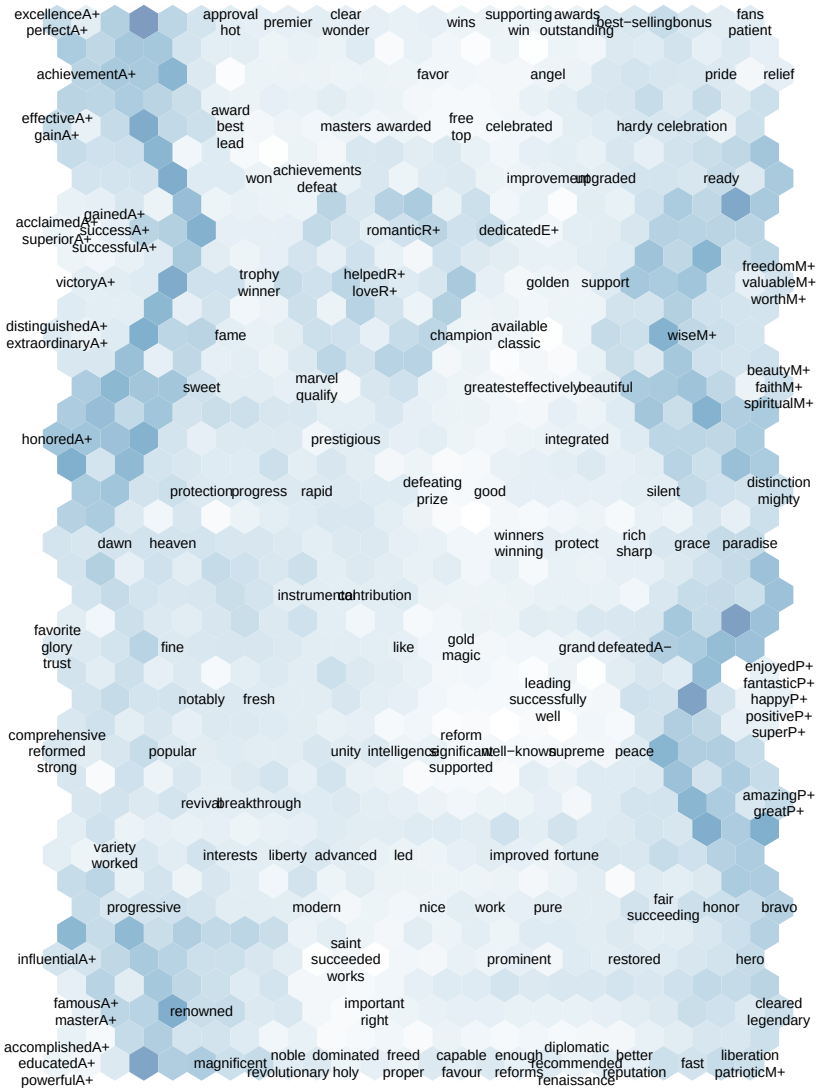

Fig. 4: A map of sentiment words based on context statistics obtained from the WikipediaA corpus. The words that belong to the PERMA lexicon are marked with a label that indicates the category (see Sec. 2 for an explanation).

useful since manual analysis of student essays is usually limited to rather small numbers of cases. In our case, a qualitative analysis of 304 essays would be a considerable human effort but still manageable. The text mining approach, on the other hand, scales up to thousands and even millions of documents.

Interesting quantitative results can be gained when large corpora are available, collected, e.g., from social media with additional profile information [20]. A lexicon-based approach can be expanded to include sentence-level analysis to take into account context effects and to improve the precision of the analysis [23]. In this paper, we consider some of the problems and solutions related to crossing language borders. One direction is to study more carefully the philosophical and practical aspects related to multilingual and multicultural studies in this area.

Our longer term goal is develop methodology for the text mining and quantitative analysis of texts in the framework of positive psychology. Substantial developments in this field have taken place recently (cf. [20]), partly based on earlier developments (cf., e.g., [24]). Our intention is to experiment with different statistical machine learning and neural-network methods and to facilitate approaches for analyzing corpora written in other languages than English.

Acknowledgments. The authors are highly grateful to Professor Martin Seligman, the Director of the Positive Psychology Center at University of Pennsylvania with his research team for making the PERMA lexicon available. T.H. acknowledges support for the latest stages of the work from the EU Commission through its European Regional Development Fund, and the program "Leverage from the EU 2007-2013".

References

1. Castellani, B., Hafferty, F.W.: Sociology and Complexity Science: A New Field of Inquiry. Springer (2009)
2. Du Bois, J.W.: Santa Barbara Corpus of Spoken American English. University of California, Santa Barbara Center for the Study of Discourse (2000)
3. Goldspink, C.: Methodological implications of complex systems approaches to sociality: Simulation as a foundation for knowledge. Journal of Artificial Societies and Social Simulation 5(1), 1–19 (2002)
4. Hämäläinen, R.P., Saarinen, E.: Systems intelligence – the way forward? a note on Ackoff's "why few organizations adopt systems thinking". Systems Research and Behavioral Science 5(6), 821–825 (2008)
5. Hansen, L.K., Arvidsson, A., Nielsen, F.Å., Colleoni, E., Etter, M.: Good friends, bad news - affect and virality in twitter. In: The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet 2011). pp. 34–43 (2011)
6. Honkela, T., Pulkki, V., Kohonen, T.: Contextual relations of words in Grimm tales, analyzed by self-organizing map. In: Fogelman-Soulié, F., Gallinari, P. (eds.) Proc. of ICANN'95. vol. II, pp. 3–7. EC2, Nanterre, France (1995)
7. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 168–177. ACM (2004)

8. Hyman, P.: In the year of disruptive education. Communications of the ACM 55(12), 20–22 (2012)
9. Janasik, N., Honkela, T., Bruun, H.: Text mining in qualitative research application of an unsupervised learning method. Organizational Research Methods 12(3), 436–460 (2009)
10. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit. vol. 5 (2005)
11. Kohonen, T.: Self-Organizing maps. Springer, Heidelberg (2001)
12. Koskenniemi, K.: A general computational model for word-form recognition and production. In: Proceedings of the 10th international conference on Computational linguistics. pp. 178–181. Association for Computational Linguistics (1984)
13. Lindén, K., Silfverberg, M., Pirinen, T.: HFST tools for morphology–an efficient open-source package for construction of morphological analyzers. In: State of the Art in Computational Morphology, pp. 28–47. Springer (2009)
14. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of LREC'10. ELRA, Valletta, Malta (2010)
15. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2(1-2), 1–135 (2008)
16. Ritter, H., Kohonen, T.: Self-organizing semantic maps. Biological Cybernetics 61(4), 241–254 (1989)
17. Saarinen, E.: Life-philosophical lecturing as a systems-intelligent technology of the self. In: The XXIII World Congress of Philosophy, Athens, Greece (2013)
18. Saarinen, E., Lehti, T.: Inducing mindfulness through life-philosophical lecturing, p. to appear. Wiley (2014)
19. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Lucas, R.E., Agrawal, M., Park, G.J., Lakshmikanth, S.K., Jha, S., Seligman, M.E.P., Ungar, L.H.: Characterizing geographic variation in well-being using tweets. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM) (2013)
20. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one 8(9), e73791 (2013)
21. Seligman, M.E.: Flourish: A visionary new understanding of happiness and well-being. Free Press, New York, NY (2011)
22. Seligman, M.E., Csikszentmihalyi, M.: Positive psychology: An introduction. American Psychologist pp. 5–14 (2000)
23. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1631–1642. Association for Computational Linguistics, Stroudsburg, PA (2013)
24. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology 29(1), 24–54 (2010)
25. Turney, P.D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 417–424. Association for Computational Linguistics, Stroudsburg, PA, USA (2002)