Sentiment Analysis Using Computer-Assisted Text Analysis Tools

Saroj S. Date(✉), Kiran V. Sonkamble, and Sachin N. Deshmukh

Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, MS, India

Abstract. Recently the use of computerized text analysis tools to assess an individual's linguistic, emotional and psychological characteristics has exploded in the field of empirical psychology. As a result, information about what people convey through their words can be swiftly and reliably extracted and analyzed. The key purpose of this research work is to analyze text data to assess linguistic and emotional characteristics with the help of computer-assisted text analysis tools. The analysis employed widely available text and sentiment analysis tools, Empath and LIWC. As text data, children's storybook reviews were analyzed in this work. These reviews are written by the children for the children. Empath and LIWC tools helped to measure the reviewer's sentiment, analytical ability and cognition level. Finally, by calculating the Pearson correlation coefficient for the selected variables, it is inferred that Empath shares a high correlation with LIWC.

Keywords: sentiment analysis · computer-assisted text analysis (CATA) tools · computer-aided text analysis software · text data analysis · LIWC · Empath · content analysis

1 Introduction

An emerging area of research in computational social science and human-computer interaction makes use of technologies to interpret the emotions and sentiments expressed in natural language. For example: since more than a couple of decades, scholarly writing analysis has been the subject of active cognitive research. It uses techniques like time-consuming traditional methods to automated text analysis tools. To extract high-quality information from text, such as cognition level, communication, authenticity, tone, analytical ability etc. the automated techniques combine linguistic, statistical and machine learning methods. For automated text analysis, researchers collect the data from common sources like social media, product reviews, news articles, blogs, etc.

As social media users have access to ever larger and more diverse data from the Internet, it becomes significant to scale our ability to conduct such analyses with breadth and accuracy. In this paper, reviews written for children's storybooks are collected to perform sentiment analysis tasks. These reviews are obtained from a website and written by children of different age groups. The collected reviews are processed and analyzed with the help of computer-assisted text analysis (CATA) tools: LIWC and Empath.

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 671–679, 2023.
https://doi.org/10.2991/978-94-6463-136-4_58

1.1 What is CATA?

CATA stands for **C**omputer-**A**ided **T**ext **A**nalysis or **C**omputer-**A**ssisted **T**ext **A**nalysis.

CATA is prominently useful for those researchers who aim to capture the emotions, cognition and beliefs of individuals as reflected in their narratives and written texts. The main entities of the CATA tools are some kind of internal dictionaries. However, researchers may develop their own dictionaries for text analysis.

There are different approaches for computerized text analysis tasks. Some of the approaches are mentioned as follows. It might be beneficial to present these, before describing the findings in the literature.

- One of the most basic approaches for text analysis is the traditional approach. It involves judgment-based content analysis by human beings. Human judges, such as subject matter experts, study and categorize a sample of textual data based on content similarities.
- Basic computer-based automation techniques are largely used to identify text data. These techniques simply count the word frequencies and categorize the data. It may also display the results in visual forms like word clouds.
- Another popular computerized technique to find out the sentiment of text data is the use of existing data dictionaries. In addition, researchers can design their own dictionaries to measure certain constructs and employ them for content analysis. It is similar to a keyword search method in which the researcher identifies all relevant phrases related to a construct and then searches the texts for these words. Publicly some dictionaries are available for sentiment classification like LIWC and DICTION.
- Advanced Natural Language Processing (NLP) tools and techniques (like latent Dirichlet analysis (LDA) and latent semantic analysis (LSA) can be employed to identify the language constructs that exist in a corpus. These techniques may be combined with data dictionaries and basic automation.

To analyze text data, some software are available like Nvivo, Linguistic Inquiry Word Count (LIWC), Atlas.ti, DICTION, Cat Scanner, MonoConc Pro, General Inquirer (GI), Empath, Wordstat, Leximancer, Textpak, Textual Analysis Computing Tools (TACT), Automap.

In this paper, to analyze the text data, LIWC and Empath tools are used. Empath is one of the modern text analysis tools. It enables researchers to build and test new lexical categories on demand by combining machine learning approaches and crowdsourcing.

LIWC is a good software to analyze text corpus by counting words in lexical categories. It analyzes social, cognitive, emotional, and other psychological dimensions within the written text. It has significant features like fast data processing, easy interpretation of the results and extensively validated dictionary.

2 Literature Survey

The significant research work carried out in the domain of sentiment analysis which uses computer-assisted text analysis (CATA) tools is presented in this section. Researchers of this domain used CATA for emotion/sentiment analysis. As per the survey, there is a

considerable increase in the usage of this type of software for sentiment analysis as well as evaluating linguistic and psychological properties of human language.

Over the last two decades, Emotion/Sentiment Analysis from text has been regarded as a demanding and interesting task. Vaibhav Tripathi et al. conducted an extensive survey on the computational analysis of emotions. They compiled a list of the various methodologies, datasets, and resources for sentiment analysis [1].

Zachary Dau reported outcomes on a computational analysis of Twitter activity by politicians. He found a significant increase in Twitter activity throughout the pandemic [2]. In the subject of Human Resource Management, Emily D. Campion and Michael A. Campion employed computer-assisted text analysis [3].

Von Selasinsky et al. used computer-aided text analysis to investigate crowdfunding success factors. They examined textual information from video subtitles as well as project titles and descriptions. They discovered that including subtitle information enhances the variation explained by the respective models and, as a result, their predictive potential for financing success [4]. Bumsoo Kim used computer-assisted text analysis to determine which social grooming characteristics diminish incivility among social media users when discussing or publishing on the COVID-19 situation in South Korea. According to the findings, the size of one's social network is a negative predictor of civility [5].

Maverick Ferreira et al. demonstrated a method for automatically classifying the content of messages in online discussions. To accomplish this task, he presented a method based on a mix of classic text mining characteristics and word counts retrieved using proven linguistic frames [6]. Shihab Elbagir and Jing Yang classified sentiments conveyed in Twitter data using the Valence Aware Dictionary for sEntimentReasoner (VADER) [7].

For sentiment analysis of the most "Fanned" Facebook Pages, Alan R. Pella employed the most widely explored method, LIWC [8]. Saifuddin Ahmed et al. studied whether online protest activities have the same emotional underpinnings as offline protest actions for sustaining and nourishing a social movement, and how these emotions alter across different stages of the social movement [9]. Ryan L. Boyd highlighted how language may reveal profound insights into the minds of others using well-established and straightforward psychometric approaches [10].

3 Methodology

In academic research, text data analysis using various approaches has a long history. The primary goal of this research work is to analyze text data with the help of modern computer-assisted text analysis tools. As written in Sect. 1, the tools under consideration are LIWC and Empath. Empath and LIWC analyses are driven by dictionary-based word counts. Empath works on a broader range of categories and may build and validate new categories on demand using unsupervised language modeling whereas LIWC provides a highly validated dictionary for analysis. After successful analysis, the obtained results are discussed in Sect. 4 of this paper. To accomplish above mentioned task, the domain of scholarly writing analysis is considered. One of the major characteristics of scholarly writing is that it should be written in concise statements and should show an understanding of the topic.

For this work, children's storybooks reviews were collected from a website. These are book reviews by children for children. Following this, the reviews were processed by custom-written Python scripts that transformed the text into a form that could be analyzed using the Empath and LIWC. Empath and LIWC software produces hundreds of measures as a result for analysis. For this study, to identify the sentiments of reviews, we limited the analysis to three variables of Empath and ten variables of LIWC. The selected Empath and LIWC category score values are processed and compared for performance analysis.

4 Results

The analysis of the results of two CATA tools are presented in this section. The storybook reviews, which are considered for experiment, are written by children of different age groups for children. The age groups are 8 to 10 Years, 10 to 12 years and 12 years and above. We labeled it as Group A, Group B and Group C respectively. Analysis on different parameters is done group-wise.

4.1 Review Text Statistics

Some of the sample reviews from each age group are given in Table 1. All of the reviews have been analyzed group-wise to find out the statistics like average word count, minimum and maximum word count and median value.

Computerized text analyses have been used to find summary of these text reviews. It is presented in Table 2. For Group A, reviews ranged from 68–155 words, with an average of 118 words per review. Group B review sentences ranged from 59–516 word with an average of 153 words per reviews. For Group C, review sentence range is 129–409 with an average of 255. As depicted in the table, it has been analyzed that the review writing characteristics have improved linearly as per the growing age group.

LIWC analysis has been carried out to find the average word count. It is shown in Fig. 1. It has been analyzed that the average word count increases as children grow up.

Also, the word clouds are generated by CATA software to see the vocabulary of the children writing. As per the analysis, top five word of Group A are: 'book', 'uncle', 'read', 'story' and 'polly'. While Group B children have used top five words as 'book', 'characters', 'story', 'michael' and 'recommend'. 'book', 'read', 'raju', 'said' and 'years' are the top 5 words by group C reviews. All these word clouds are shown in Fig. 2. From figure it can be analyzed that vocabulary of Group B and C children is better than Group A.

4.2 LIWC Output Variable Analysis

LIWC text analysis software generates hundreds of output variables for every analyzed text file. It also displays the scores for four summary language variables called as analytical ability, clout, authenticity and emotional tone. For children book review analyses, we have considered analytical ability and cognition level variable values. The process of learning information and understanding through thought, experience, and the senses is referred to as cognition. Figure 3 depicts the rising scores values of analytical ability and cognition variable with relation to the age of the children.

Table 1. Sample book reviews from different age groups

Reviewer Group	Book Title	Sample Book Review
Group A	The Snow Queen	It is a story of a boy named Kay and a girl called Grenda. Kay was kidnapped by a wicked Snow Queen and she takes Kay to her ice castle in the far north. Brenda came to rescue Kay and had lots of exciting adventures on the way and met lots of interesting people..I really enjoyed this book and it took me two days to read it. I would tell my friends that I read The Snow Queen and say it was a very good book.
Group B	The Dreamsnatcher	This 'fast paced and full of charm' action-packed thriller is an outstanding debut from Abi Elphinstone. It is packed with suspense, adventure and once you have picked up the book there is no letting go of it. Moll is a brave young child who is determined to fight off the Dreamsnatcher and with the help of Gryff, a wildcat always by Moll's side, she can do what the Oracle Bones have foretold. I could not choose a favourite part as I enjoyed the whole book greatly. I cannot wait for the sequel that will be coming out in 2016. "A superb, brilliant debut," Hugo Ding.
Group C	The Quietness	The Quietness is set in nineteenth century London. Each chapter is written from the perspective of the two main characters, Queenie and Ellen. They have different lifestyles. Queenie lives in poverty at home with her large family but circumstances lead her to run away to find a good paying job. In contrast, Ellen lives at home and her family is wealthy. They soon meet and together they form a strong bond where they then find out a secret that changes both of the girls' lives forever. I found this book interesting and gripping to read and wouldn't put the book down after the unexpected plot twist. It should me a lot about friendship and trust. This book deserves five stars and I would definitely recommend this book to girls around fourteen years old. It is good for readers who love a bit of drama and it made me actually want to read in my free time. This is a YA novel and is recommended for readers aged 14 and above.

Table 2. Summary information of book review

	Group A	Group B	Group C
Average word count	118	153	255
Minimum	68	59	129
Maximum	155	516	409
Median	110	124	267

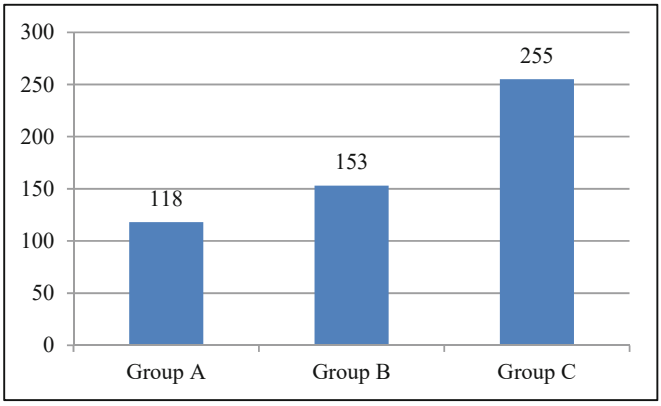

Fig. 1. Computerized analysis of reviews- average word count

4.3 Empath vs LIWC Analysis

Some of the common variables from both CATA tools are evaluated to see how closely Empath and LIWC categories are correlated. These variables are communication, positive-emotion and negative_emotion. Communication refers to exchange of information by speaking, writing, or using some other medium. A positive emotion is an emotional reaction that is intended to convey a pleasant feeling and negative emotion convey unpleasant feeling. The score values for these communications, positive_emotion and negative_emotion variables are shown in Fig. 4.

Group-wise sentiment analysis score values are shown in Table 2. These are being used to calculate the Pearson correlation coefficient (PCC). It is the most widely used measure in research field to determine the strength and direction of the relation between two variables. It is a value that ranges from -1 to 1 and represents linear correlation.With a value of -1 signifying a total linear correlation that is negative, 0 signifying no correlation, and $+1$ denoting a total linear correlation that is positive.

In Table 3, Group (ABC) denotes the calculated average scores of Groups A, B and C. After closely analyzing the average scores of the common variables, it can be inferred that selected Empath's categories are highly correlated with the respective categories of LIWC's with average PCCs of $r = 0.909$. It implies that Empath's data-driven word counts are quite close to an extensively validated LIWC dictionary.

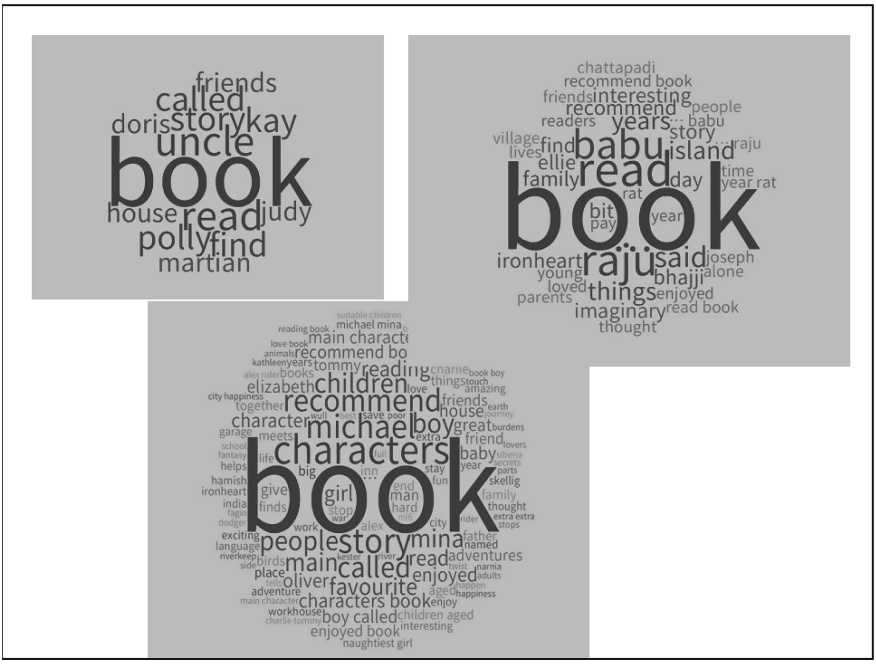

Fig. 2. Word clouds generated by CATA tools

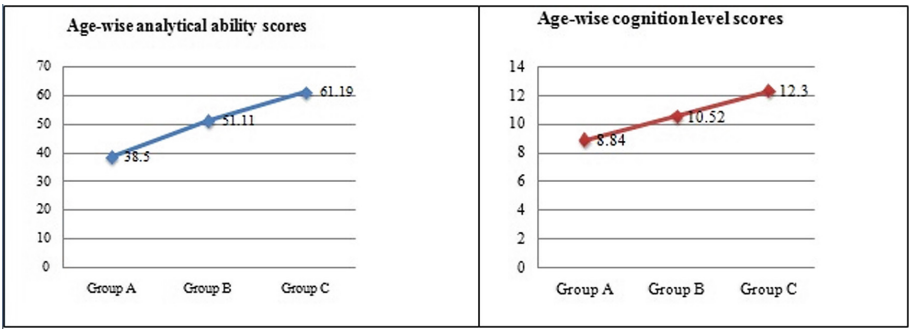

Fig. 3. LIWC generated analytical ability and cognition scores

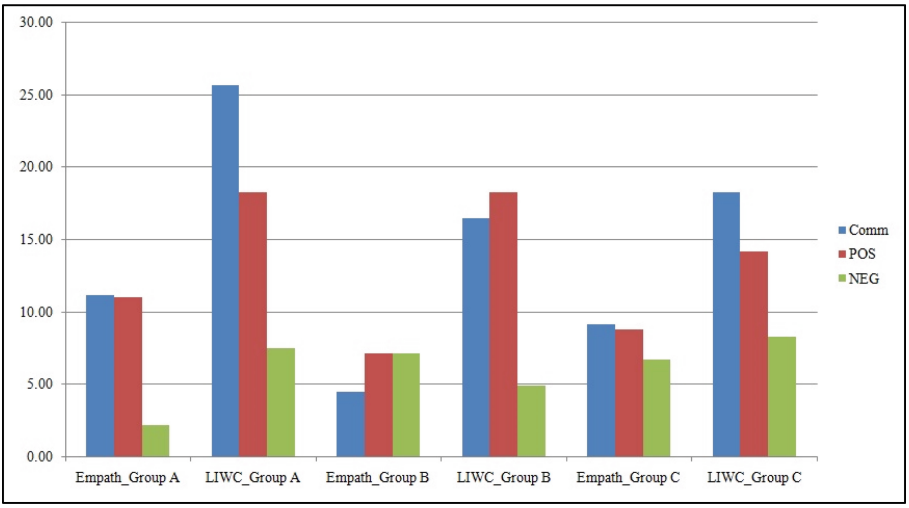

Fig. 4. Group-wise Sentiment Analysis using Empath and LIWC tools

Table 3. CATA scores for selected variables

Variable	Group A		Group B		Group C		Group(ABC)	
	Empath	LIWC	Empath	LIWC	Empath	LIWC	Empath	LIWC
Communication	11.12	25.68	4.48	16.5	9.11	18.3	8.24	20.16
Pos_emo	10.99	18.28	7.14	18.3	8.81	14.2	8.98	16.93
Neg_emo	2.16	7.46	7.10	4.9	6.69	8.3	5.32	6.89

5 Conclusion

In this paper, we presented the use of Computer-Assisted Text Analysis tools to identify natural language constructs in order to assess the linguistic and emotional characteristics of text data. To analyze storybook reviews, Empath and LIWC tools were used. From the analysis, it is scientifically proved that children's scholarly writing improves linearly with age. Through experimental work, it is concluded that Empath has a high correlation with LIWC software. However, the overall observation is that, while Empath covers a broader set of categories than LIWC, it does not include similar kinds of LIWC variables such as Analytic, Clout, Authentic, and Tone. These are LIWC's highly condensed summary variables. Another noting is that, Empath can generate and validate new categories with a few seed words. LIWC's dictionaries are created and validated rigorously using semi-automatic approaches.

This work has some limitations. The focus is on analyzing storybook reviews, which means that the findings may not be applicable beyond this domain. Nonetheless, we would expect similar results for the same domain. Future researchers can extend this work by comparing and contrasting the results obtained with other text analysis software.

They can also assess other aspects of LIWC and Empath besides analytical ability, cognition, communication, positive emotion and negative emotion.

References

1. Tripathi, V., Joshi, A., Bhattacharyya, P.: Emotion analysis from text: A survey. Center for Indian Language Technology Surveys 11(8), 66–69 (2016).
2. Dau, Z.: A Computational Analysis of the Coronavirus Pandemic Response of Tri-State Area Politicians on Twitter. In Corpus Linguistics, (2021).
3. Campion, E. D., Campion, M. A.: Using Computer-assisted Text Analysis (CATA) to Inform Employment Decisions: Approaches, Software, and Findings. Research in Personnel and Human Resources Management, (2020).
4. Von Selasinsky, C., Isaak, A. J.: It's all in the (Sub-) title? Expanding Signal Evaluation in Crowdfunding Research. arXiv preprint arXiv:2010.14389, (2020).
5. Kim, B.: Effects of social grooming on incivility in COVID-19. Cyberpsychology, Behavior, and Social Networking 23(8), 519–525 (2020).
6. Ferreira, M., Rolim, V., Mello, R. F., Lins, R. D., Chen, G., Gašević, D.: Towards automatic content analysis of social presence in transcripts of online discussions. In Proceedings of the tenth international conference on learning analytics & knowledge, pp. 141–150 (2020).
7. Elbagir, S., Yang, J.: Twitter sentiment analysis using natural language toolkit and VADER sentiment. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 122, p. 16), (2019).
8. Peslak, A.: Facebook Fanatics: A Linguistic and Sentiment Analysis of the Most "Fanned" Facebook Pages. Journal of Information Systems Applied Research, 11(1), 23, (2018).
9. Ahmed, S., Jaidka, K., Cho, J.: Tweeting India's Nirbhaya protest: A study of emotional dynamics in an online social movement. Social Movement Studies 16(4), 447–465(2017).
10. Boyd, R. L.: Psychological text analysis in the digital humanities. In Data analytics in digital humanities (pp. 161–189). Springer, Cham (2017).