https://doi.org/10.36315/2022inpact105
Psychological Applications and Trends 2022

LINKING MODAL AND AMODAL REPRESENTATIONS THROUGH LANGUAGE COMPUTATIONAL MODELS

Diego Iglesias, Miguel A. Sorrel, & Ricardo Olmos
Department of Social Psychology and Methodology, Universidad Autónoma de Madrid (Spain)

Abstract

Language computational models such as Latent Semantic Analysis (LSA) has been criticized for not having direct contact with the real world. However, recent findings have shown the ability of the LSA to capture embodied features such as words' emotional content. In the present study we tested whether LSA can predict the emotions contained in short written texts such as tweets. It was found that a multiple logistic regression model receiving as input LSA information classified correctly 73,9% of the tweets analyzed according to the emotional content. These results provide additional evidence underlying the representative power of abstract symbols and showing the link between modal representations (emotional) and amodal representations (abstract symbols) through the LSA.

Keywords: *Latent semantic analysis (LSA), computational language, embodiment, representation, tweets.*

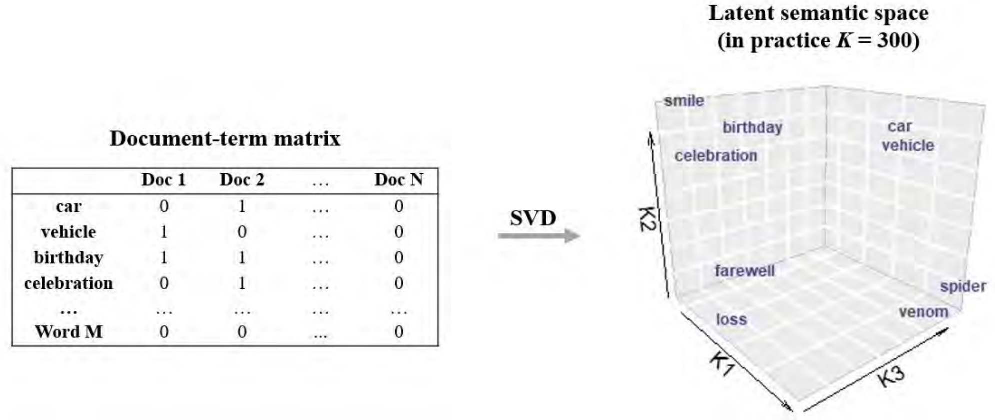
1. Introduction

Latent Semantic Analysis (LSA) is a language computational model that represents human semantics in a formal metric (Landauer & Dumais, 1997). Processing a large corpus of texts, LSA reflects the semantic relations within words creating a semantic space where words are represented as vectors (Deerwester et al., 1990).

The analysis begins segmenting the linguistic corpus into words and documents (i.e. paragraphs) and storing this information in a document-term matrix. Then, applying a dimensionality reduction technique such as the singular value decomposition (SVD), LSA searches for the number K of independent components that best resumes the information of the matrix. Thus, it creates the K dimensional semantic space where words are represented as vectors with coordinates on each of the K dimensions retained (Deerwester et al., 1990) (Figure 1).

The number of dimensions retained are those with which LSA obtains the best results simulating the human behavior (e.g. in a vocabulary task such as *TOEFL*). After a large number of experiments, 300 seems to be the number of dimensions that best represents human semantics (See Landauer et al., 1998 for a more extended explanation).

Figure 1. Latent semantic space generation. Note that this figure shows a simplified representation in three dimensions (K = 3).

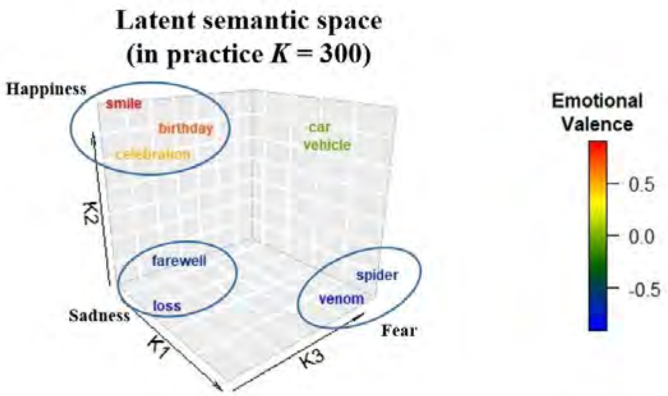


p-ISSN: 2184-2205 e-ISSN: 2184-3414 ISBN: 978-989-53614-1-0 © 2022

LSA has been criticized by the embodied theories of cognition and language (Barsalou, 1999; De Vega, 2005; Glenberg & Robertson, 2000). These approaches state that any mental process, such as language comprehension, must be grounded in the real world. They claim that LSA is unable to make contact with reality because it only processes abstract symbols without linking them to their real world referents. In response to these critiques, Landauer (1999) argues that the abstract symbols with which LSA operates are richer in information. In the end, the way humans use words like "spider" or "venom" reflects what they have learned from their real-word experience. The symbol interdependency hypothesis (Lowerse, 2011, 2018) also underlines the representative power of abstract symbols. This hypothesis proposes that corporeal properties of the real world are encoded in both modal representations (visual, auditory, emotional, etc.) and amodal representations (abstract symbols).

Recent studies (Martínez-Huertas et al., 2021) found that neural network models receiving as input LSA information were able to predict the emotional content of a sample of words. These results emphasize the idea that abstract symbols processed by computational models capture features grounded in human's world (Figure 2).

Figure 2. Amodal representations encode embodied features such as the emotionality (i.e. emotional category and valence) of words.



Based on these ideas, in the present study we test whether the LSA can detect the emotional content of short written texts such as tweets. Considering the recent findings (Martínez-Huertas et al., 2021), we expect LSA to perform properly. Thus, this study would provide new evidence in accordance with the perspectives that emphasize the representative power of symbols.

2. Method

2.1. Data sampling and vectorization in the latent semantic space

Twitter is one of the most widely used social networks where users share their life experience by short written texts such as tweets (Twitter, Inc, 2021). Using the advanced search filters from Twitter we collected all tweets including the words "he sentido…" ("I felt…") published from July 26, 2019 to March 30, 2020 in Spain. We considered that texts including this expression could contain emotional features expressing positive or negative affect. After data cleaning (i.e. removing incomprehensible tweets), the final sample of tweets was 522.

We used Gallito Studio software (Jorge-Botana et al., 2013) to generate the semantic space (K = 300) and vectorize the tweets. The space was created from a journalistic corpus composed of 150.802 documents and 23.835 words that came from articles written in 2019 in the Spanish newspapers El País and El Mundo.

2.2. Neural network models, human criterion and logistic regression

We connected to GallitoAPI (an API that allows to use Gallito Studio software online) and used the neural network models trained to propagate the emotionality of words using the latent semantic space information (Martínez-Huertas et al., 2021). In the present study, the neural network models receive as input the LSA coordinates of each tweet and give as output the four independent variables: degree of happiness, sadness, fear and anger contained on the text on a scale from 0 to 1.

To define the dependent variable (human criterion), the first author evaluated and classified the N = 522 tweets according to their emotional content in positive or negative valence. A total of 192 tweets were classified as positive and 330 as negative ($n_{positive}$= 192; $n_{negative}$= 330). Testing the reliability of the evaluation, we found total agreement (*Cohen's kappa* = 1,00) on a subsample of 100 tweets between the first author and two expert judges in Psychology (a 61-year-old man and a 54-year-old woman, professors at the Faculty of Psychology of the Universidad Autónoma de Madrid).

Then, we fitted a multiple logistic regression model to classify the tweets as *positive* = 0 or *negative* = 1 (dependent variable) based on the LSA information (the four independent variables: *degree of happiness*, *sadness*, *fear* and *anger*). The cutoff point to classify a tweet as negative was set at 0.5.

The whole analysis (connection with GallitoAPI and statistical tests) was conducted using R software (R Core Team, 2021).

3. Results

The null model (i.e. the simplest model), which does not include any independent variable classifies all tweets as negative ($-2LL_0$ = 686,726; 63,2% of tweets correctly classified).

The logistic regression model with the four independent variables fitted the data significantly better than the null model ($G^2_{(4)}$ = 140,678, p<0,001) and explains 32% of the dependent variable variance ($R^2_{Nagelkerke}$ = 0,32). As shown in Table 1, 73,9% of the tweets were correctly classified according to the human criterion.

As can be noted from the table, only the *degree of happiness* (*logit B* = -0,912; p<0,001) and *sadness* (*logit B* = 1,068; p<0,001) contributed to explain the observed variance of the dependent variable (i.e. to improve predictions). Considering how the dependent variable has been operationalized (*positive* = 0, *negative* = 1), the positive sign of *degree of sadness* (*logit B* = 1,068) indicates that augmenting the amount of sadness of a tweet would increase the probability of being classified as negative. Thus, augmenting the *degree of sadness* one unit would increase 191% the *odds* of being classified as negative ($Exp(B)_{degree\_sadness}$ = 2,91). In contrast, augmenting the *degree of happiness* (*logit B* = -0,912) one unit would decrease 60% the odds of being classified as negative ($Exp(B)_{degree\_happiness}$ = 0,40).

Table 1. Logistc regression results: a) classification and b) coefficients.

a)	**Model prediction**		**% tweets correctly classified**	
First author evaluation	**Positive**	**Negative**		
Positive	102	90	Positives = 53,1%	
Negative	46	284	Negatives = 86,1%	
			Total = 73,9%	
b) Regression coefficents				
	Logit B	***p value***	**Exp (*B*)**	**IC 95%**
Degree of happiness	-0,912	0,000	0,402	0,296 – 0,545
Degree of sadness	1,068	0,000	2,909	1,723 – 4,911
Degre of fear	-0,012	0,969	0,988	0,551 – 1,772
Degree of anger	0,574	0,067	1,776	0,960 – 3,287

4. Discussion and conclusions

The results obtained in the present study show that the evaluation of the emotional valence (i.e. positive or negative) of tweets performed by neural networks model receiving as input LSA information concurs with human criterion. Using the output generated by the neural networks as independent variables (*degree of happiness, sadness, fear* and *anger*), the multiple logistic regression model classified correctly 73,9% of the tweets, which is a considerably high hit rate. Moreover, the estimated regression coefficients for the statistically significant variables (*degree of happiness and sadness*) are consistent with theory. Augmenting the sadness (i.e. negative affect) in a tweet increases the probability of being classified as negative. However, increasing the happiness (i.e. positive affect) decreases the probability of being classified as negative.

It is important to note that the neural networks used in this study were trained to predict the emotionality of words instead of short written texts such as tweets. In addition, the human criterion that defined the dependent variable was based on the judgement of a single participant (the first author). These limitations can be overcome in future research introducing software tools (e.g. R software) to automate the data collection process. Large amounts of data (i.e. tweets) will make it possible to train neural network models to predict the emotionality of tweets and also to improve the criterion validity (e.g. holding-out subsamples of tweets to be evaluated by humans).

Considering the performance of the logistic regression model (i.e. considerably high hit rate and coefficients consistent with theory), the results obtained in the present study are consistent with the ideas of Landauer (1999), Lowerse (2011, 2018) and Martínez-Huertas et al. (2021) and emphasize the representative power of symbols showing that amodal representations processed by LSA encode embodied features such as the emotionality of the tweets.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577-660.

De Vega, M. (2005). Lenguaje, corporeidad y cerebro: Una revisión crítica. *Revista Signos*, *38*(58).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41(6)*, 391-407.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, *43*(3), 379-401.

Jorge-Botana, G., Olmos, R., & Barroso, A. (2013). Gallito 2.0: A natural language processing tool to support research on discourse. In *Proceedings of the 23th Annual Meeting of the Society for Text and Discourse*, Valencia, España, 16-18 Julio.

Landauer, T. K. (1999). Latent semantic analysis (LSA), a disembodied learning machine, acquires human word meaning vicariously from language alone. *Behavioral and brain sciences*, *22*(4), 624-625.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259-284.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273-302.

Louwerse, M. M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, *10*(3), 573-589.

Martínez-Huertas, J. Á., Jorge-Botana, G., Luzón, J. M., & Olmos, R. (2021). Redundancy, isomorphism, and propagative mechanisms between emotional and amodal representations of words: A computational study. *Memory & Cognition*, *49*(2), 219-234.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Twitter, Inc. (2021). *Twitter*. Microsoft Store. https://www.microsoft.com/es-es/p/twitter/9wzdncrfj140?activetab=pivot:overviewtab