Sound Symbolism in Automatic Emotion Recognition and Sentiment Analysis

Alexander J. Kilpatrick [1]

[1] *Nagoya University of Commerce and Business, Nisshin, Japan.*

Abstract

This report documents the construction and output of extreme gradient boosted algorithms that were trained using the phonemes that make up American English words to identify how different sounds express emotion and sentiment. The data comprised of two corpora that consist of words that have been assigned scores according to how they reflect certain emotions and sentiments. The models are trained only on the phonemes that make up each word. This is a unique approach to automatic emotion recognition and sentiment analysis which typically does not consider individual phonemes. In addition to the boosted algorithms, linear regression is used to examine the relationships between word length, and emotions and sentiments.

Keywords

Sound Symbolism; Automatic Emotion Recognition; Automatic Sentiment Analysis; XGBoost; Extreme Gradient Boosting; Artificial Intelligence; Human-AI Interaction.

1. Introduction

The principle of the arbitrariness of the sign [1] is a foundational concept in linguistics. It posits that there is no inherent or logical connection between the sound of a word and its meaning. In other words, linguistic signs are considered to be arbitrary, with their meanings assigned by convention rather than by any intrinsic relationship between sound and sense. In recent years, there has been a growing interest in exploring sound symbolism, challenging the notion of the arbitrariness of the sign (e.g., [2], [3], [4], [5]). These studies have revealed that concepts such as size and shape exhibit stochastic relationships with sounds. Furthermore, many of these relationships are found to be consistent across different languages. For instance, the *mil/mal effect*, which observes that vowels like /i/ and /a/ are more frequently used in the names of small and large referents respectively, and the *kiki/bouba effect*, which notes that spiky shapes are often associated with sounds resembling *kiki*, while rounded shapes are often associated with sounds resembling *bouba*, have been found to hold true in numerous languages around the world (e.g., [2], [4]). Emotional sound symbolism refers to the phenomenon where the sounds of words are associated with, and convey, specific emotional or affective qualities. In this linguistic concept, certain phonemes or combinations of sounds are thought to evoke or symbolize particular emotional states or feelings. In a cross-linguistic study of five languages [6], researchers observed a pattern whereby phonemes at the beginnings of words predict emotional valence most strongly and that phonemes associated with negative valence were uttered more quickly, drawing parallels between emotional sound symbolism and alarm calls in the animal kingdom. Important to the present study, they observed that in English, phonemes like /ʌ/, /d/, /ɪ/, and /oʊ/ were associated with negative valence while phonemes like /tʃ/, /ɛ/, and /p/ were associated with positive valence.

Automatic emotion recognition and sentiment analysis is a subfield of natural language processing which endeavors to construct algorithms that understand human emotion and sentiment in language. Sound symbolism has been largely overlooked in automatic emotion recognition and sentiment analysis although there are a few studies that explore sound symbolism through the lens of machine learning. For example, Winter and Perlman [7] constructed random forest algorithms to show a systematic size-sound relationship in English size adjectives. As with the present study, the data was engineered in a

Cognitive AI 2023,13th-15th November, 2023, Bari, Italy.
EMAIL: alexander_kilpatrick@nucba.ac.jp (A. 1)
ORCID: 0000-0003-3134-3797 (A. 1)

manner so that each sample returned a count of the number of times each phoneme occurs in each word. The outcome of this is a dataset comprised of mostly null values. Following on this method researchers constructed algorithms to classify Pokémon names according to their evolution level using sound symbolism [8]. They showed that the random forest algorithms were able to classify novel Pokémon names more accurately than Japanese university students assigned to an identical task. An issue of overfitting due to the high number of null values in the dataset was uncovered and resolved using cross-validation. In the present study, word length is found to be a significant predictor of several emotions and sentiments and this effect is taken into consideration in the design and analysis of the algorithms. Potentially related, Li et al. [9] examined the relationship between utterance length and word error rates in automatic speech recognition and speech emotion recognition. They found that shorter utterances tended to have higher word error rates likely due to a lack of contextual information.

The present report outlines the construction and output of algorithms designed to combine sound symbolism with automatic emotion recognition and sentiment analysis. Two corpora, with a combined total of almost 20,000 words, are used to train 19 algorithms that are each designed to classify samples according to specific emotions and sentiments. All algorithms return significant accuracy estimates.

2. Method

All data, files, codes, and links to a YouTube series documenting this project can be found in the following online repository: https://osf.io/brus3/?view_only=63412.

The present study uses two separate corpora to train machine learning algorithms. The first is the Glasgow Norms [10], a list of 5,500 words that have been assigned Likert scores for 9 sentiments. A full list of the sentiments in the Glasgow Norms is provided in Table 1. The second corpus is the NRC Word-Emotion Association Lexicon [11], a list of 14,000 words that have been assigned a binary score as to whether each word is associated with 10 emotions and sentiments. A full list of the emotions and sentiments in the NRC Lexicon can be found in Table 2. Words from both corpora were cross referenced with the Carnegie Mellon University Pronouncing Dictionary (CMU [12]) to obtain American English phonemes for each word. Words that did not find a match in the CMU were manually checked. Instances of mismatched spelling were corrected. All other unmatched samples were discarded.

All analyses were conducted in the R environment [13]. Word length was calculated by summing the number of phonemes in each word. No additional considerations were made for diphthongs or long vowels which were counted as single phonemes. The relationships between word length, and emotions and sentiments were analyzed using a series of regression equations, dependent variables being the average Likert scores in the Glasgow Norms and the binary classification in the NRC Lexicon; independent variables being word length. The XGBoost algorithms were constructed using the `XGBoost` [14] and `caret` [15] packages. K-fold cross-validation (K = 28) was used to avoid the overfitting issue reported in [8]. The data was split into 8 subsets (A-H) and recombined using a Latin square resulting in 28 subsets with a 3:1 training to testing split. For example, the first iteration of each model is trained using subsets A through F and tested on subsets G and H. The following results report on the aggregate of each series of 28 iterations. Combined significance was calculated using Stouffer's [16] and Fisher's [17] methods; however only Fisher's method is reported as it returned more conservative significance estimates. The algorithms for both corpora were designed to classify samples so the Likert scale values in the Glasgow Norms were assigned to binary categories using a median split. The XGBoost algorithm was found to be susceptible to distribution skew, so categories were balanced by randomly removing samples from the majority category. This had little effect on the Glasgow Norms dataset due to the median split but removed around 80% of samples in the NRC dataset because only around 10% of samples in that dataset have a value of 1 in the binary dependent variable. To increase variability, balancing was conducted after cross-validation sub-setting. To limit the influence of word length in the XGBoost models, phoneme counts were divided by word length so that features were a percentage how much a phoneme makes up each word. This resulted in a convergence issue during tuning, so α was manually adjusted and the same learning rate was applied to all models ($\alpha = 0.1$). All other hyperparameters were automatically tuned by inputting diverse hyperparameter settings into a tuning grid.

3. Results

3.1. Linear Regression and Word Length

A series of linear regression models were calculated to test the relationship between word length, and the emotions and sentiments in the Glasgow Norms (Likert score) and the NRC Lexicon (binary). Table 1 reports on the findings of the analyses conducted on the Glasgow Norms. Increased Age of Acquisition, Arousal, Size, and Valence had a significant positive correlation with word length while Concreteness, Familiarity, and Imaginability had a negative one. All significant relationships observed in the analyses conducted on the NRC Lexicon (Table 2) showed a positive correlation. These include Anger, Sadness, and Trust emotions while the Negative and Positive sentiments were also significant.

Table 1

Word Length and the Glasgow Norms.

Sentiment	F-statistic	p-value	R^2
Age of Acquisition	986.4	< 0.001	0.152
Arousal	152.7	< 0.001	0.027
Concreteness	321.5	< 0.001	0.055
Dominance	0.273	0.601	0
Familiarity	63.72	< 0.001	0.011
Gender	0.617	0.432	0
Imaginability	333.4	< 0.001	0.057
Size	578.3	< 0.001	0.095
Valence	5.928	0.015	0.001

Table 2

Word Length and the NRC Lexicon

Variable	Measure	t-value	p-value
Anger	Emotion	2.099	0.036
Anticipation	Emotion	1.835	0.067
Disgust	Emotion	-0.241	0.809
Fear	Emotion	0.252	0.801
Joy	Emotion	-0.02	0.984
Negative	Sentiment	3.571	< 0.001
Positive	Sentiment	7.572	< 0.001
Sadness	Emotion	2.467	0.014
Surprise	Emotion	0.125	0.901
Trust	Emotion	4.965	< 0.001

3.2. XGBoost Accuracy

All models constructed and tested using the Glasgow Norms achieved an accuracy greater than chance and a Fisher's combined p-value < 0.001. Table 3 reports on the aggregated accuracy and standard deviation of these models. A similar result was found in the NRC models except in the case of the Surprise algorithm ($p = 0.022$ using Fisher's method and $p = 0.021$ using Stouffer's method). The NRC models are presented in Table 4. The Glasgow Norms models did report a greater accuracy than the NRC models; however, it is important to note that these were constructed and tested on larger datasets due to the balancing outlined in 2. The accuracy was, on average, higher and variability was lower in the models constructed using the Glasgow Norms models compared to the NRC models.

Table 3
Glasgow Norms model accuracy (ACC) and standard deviation (STD) and Fisher's combined p values (*p*).

Sentiment	ACC	STD	*p*
Age of Acquisition	63%	1%	< 0.001
Arousal	58%	1%	< 0.001
Concreteness	61%	1%	< 0.001
Dominance	53%	1%	< 0.001
Familiarity	56%	1%	< 0.001
Gender	58%	1%	< 0.001
Imaginability	62%	1%	< 0.001
Size	58%	1%	< 0.001
Valence	56%	1%	< 0.001

Table 4
NRC model accuracy (ACC) and standard deviation (STD) and Fisher's combined p values (*p*).

Variable	ACC	STD	*p*
Anger	54%	2%	< 0.001
Anticipation	53%	2%	< 0.001
Disgust	55%	2%	< 0.001
Fear	54%	2%	< 0.001
Joy	53%	3%	< 0.001
Negative	54%	1%	< 0.001
Positive	54%	2%	< 0.001
Sadness	54%	2%	< 0.001
Surprise	51%	3%	0.022
Trust	52%	2%	< 0.001

3.3. XGBoost Feature Importance

Tables 5 and 6 report the 15 most important features for the Glasgow Norms and NRC models respectively. Certain features are consistently important across models despite measuring different emotions. Features with high feature importance across models include voiceless plosives (/t/ and /k/), the alveolar nasal (/n/), approximant consonants (/ɹ/ and /l/), the alveolar fricative (/s/), and the open-mid back vowel (/ʌ/) which appears particularly important, but it should be noted that this is also the most common phoneme in American English according to the CMU.

Table 5
The most important phonemes (IPA) and their feature importance (IMP) in the Glasgow Norms models.

Age of Acquisition		Arousal		Concreteness		Dominance		Familiarity		Gender		Imaginability		Size		Valence	
IPA	IMP	IPA	IMP	IPA	IMP	IPA	IMP	IPA	IMP	IPA	IMP	IPA	IMP	IPA	IMP	IPA	IMP
ʌ	100	l	93	ɪ	98	ɹ	91	n	90	k	93	ʌ	96	ʌ	97	ɹ	99
t	60	ʌ	89	ʌ	89	s	89	l	89	l	87	ɪ	91	t	85	n	84
s	51	t	88	l	71	l	89	t	85	t	87	l	85	ɹ	72	l	83
ɪ	42	k	88	p	70	t	81	ʌ	81	ɹ	86	k	77	k	72	k	80
l	41	ɹ	84	t	69	k	76	ɹ	79	ɪ	75	p	72	s	67	t	75
ɹ	40	d	79	k	63	ʌ	73	k	79	s	70	t	72	l	61	d	72
k	37	ɪ	77	ɹ	63	d	69	s	77	n	69	ɛ	69	p	61	s	71
p	36	s	75	ɛ	62	n	65	p	71	d	67	ɹ	69	d	51	ʌ	71
n	31	p	72	n	52	p	61	d	68	p	63	s	64	m	50	ɪ	56
d	28	n	72	i	51	m	61	ɪ	58	ʌ	60	n	57	n	48	m	54
ɛ	27	m	57	s	50	i	55	m	50	ɛ	58	ɝ	54	ɪ	42	p	53
m	27	b	52	b	46	ɝ	54	ɝ	49	i	55	d	54	ɝ	40	ɝ	50
ɝ	26	ɛ	50	d	45	ɪ	54	f	49	f	50	i	49	i	37	aɪ	48
i	23	ɝ	50	ɝ	45	b	51	i	45	b	50	b	46	b	32	i	48
b	22	æ	47	eɪ	40	æ	50	b	44	eɪ	47	f	43	ɛ	32	æ	47

Table 6
The most important phonemes (IPA) and their feature importance (IMP) in the NRC models.

Anger		Anticipation		Disgust		Fear		Joy		Negative		Positive		Sadness		Surprise		Trust	
IPA	**IMP**	**IPA**	**IMP**	**IPA**	**IMP**	**IPA**	**IMP**	**IPA**	**IMP**	**IPA**	**IMP**	**IPA**	**IMP**	**IPA**	**IMP**	**IPA**	**IMP**	**IPA**	**IMP**
s	85	ʌ	89	ʌ	95	ʌ	93	ʌ	91	ʌ	96	ʌ	99	ʌ	88	ʌ	89	ʌ	92
t	85	t	87	s	84	n	86	n	81	n	85	n	77	l	88	t	87	n	84
ʌ	84	n	82	ɪ	80	t	84	t	79	ɹ	82	t	75	ɪ	86	s	78	t	81
n	78	ɹ	75	n	79	l	75	s	76	d	81	ɪ	71	s	85	n	75	k	79
l	76	ɪ	72	l	76	ɹ	74	l	72	s	80	s	70	t	81	ɹ	71	ɹ	77
ɪ	75	s	71	t	75	s	68	ɪ	72	ɪ	79	ɹ	65	ɹ	76	ɪ	67	l	74
d	74	l	68	d	69	k	65	ɹ	67	t	76	l	62	n	76	l	66	ɪ	74
ɹ	74	k	62	ɹ	69	ɪ	64	k	62	l	75	k	60	d	75	k	61	s	68
k	61	i	56	i	61	d	64	d	55	k	59	d	54	k	61	i	51	ɝ	56
ɝ	54	p	54	k	60	m	51	p	50	p	54	i	49	i	56	d	50	ɛ	56
æ	50	d	52	m	54	ɝ	50	ɛ	48	i	53	m	48	p	53	p	47	d	55
p	46	m	48	ɝ	50	p	50	i	48	m	50	ɝ	48	m	52	ɝ	45	i	53
i	46	ɝ	45	p	48	ɛ	49	m	47	b	49	ɛ	46	ɛ	46	ɛ	44	p	49
m	46	ɛ	44	æ	48	i	46	ɝ	45	ɝ	48	p	46	ɝ	46	æ	43	æ	46
b	41	ɔ	42	ɛ	40	b	46	f	44	æ	46	æ	40	æ	44	m	41	m	45

4. Discussion

All models achieved accuracy significantly greater than chance ($p < 0.001$ in all cases but one). Although further investigation is recommended, the results suggest that it is unlikely that sound symbolism in American English expresses fine-grained emotions and sentiments because the feature importance scores suggest many models are using the same features to make decisions. Rather, it seems that sound symbolism communicates emotional and sentimental weight. Consider that the Valence model in the Glasgow Norms—where high valence is positive and low valence is negative—and the Positive and Negative models in the NRC Lexicon all showed a positive correlation with word length. Positivity and negativity are sound symbolically expressed through longer words, although this is slightly stronger for positive sentiments as shown by the NRC lexicon models and the significant, but relatively weak, positive correlation in the Valence regression model.

Those sounds that have high feature importance scores across models include voiceless plosives (/t/ and /k/), the alveolar nasal (/n/), approximant consonants (/ɹ/ and /l/), the alveolar fricative (/s/) and the open-mid back vowel (/ʌ/). Most of the consistently important consonants are produced at the alveolar ridge. /ʌ/ appears to be an especially important feature across models. This observation falls in line with [6] who showed that /ʌ/ was associated with negative valence; however, few other patterns can be drawn between the that study and the present report. The high importance of /ʌ/ might also be due to a combination of its high occurrence frequency, being the most common phoneme in the CMU, and the distribution of null values in independent variables (NRC = 85%; Glasgow Norms = 88%). XGBoost algorithms are constructed using decision trees which base their decisions upon the outcomes of nodes. At each node a certain number of features are tested. /ʌ/ is the most commonly occurring phoneme in English and it will often be tested against low frequency features with null values. This issue was somewhat mitigated by dividing phoneme counts against word length, but it doesn't solve the problem entirely. Take for example Age of Acquisition Likert scores which were shown to have the strongest association with increased word length across all models, this is an unsurprising finding. However, Age of Acquisition XGBoost model feature importance scores revealed that /ʌ/ was the most important feature in that model, to a much greater degree than other models. This suggests that word length is still contributing to the models despite attempts to mitigate its influence through model tuning and data engineering. Word length could be included in the XGBoost models; however, given that length has a greater range than phoneme counts and no null values, this would likely mask weaker features [8].

That said, all models reported significant accuracy. Given that most automatic emotion recognition and sentiment analysis systems rely heavily on lexical and syntactic features, this study underscores the potential of phonemic information as an additional valuable resource for improving the accuracy of such systems, especially when dealing with emotional and sentimental aspects of language. While the current study provides valuable insights into the role of sound symbolism in sentiment analysis, future research could delve further into the interplay between phonemic features and linguistic and contextual factors to enhance the robustness and generalizability of sentiment analysis models across different languages and domains.

5. References

[1] Saussure, F. D. (1916). *Cours de linguistique générale*. Paris: Payot.

[2] Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., ... & Winter, B. (2022). The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B*, *377*(1841), 20200390.

[3] Fort, M., Lammertink, I., Peperkamp, S., Guevara-Rukoz, A., Fikkert, P., & Tsuji, S. (2018). Symbouki: a meta-analysis on the emergence of sound symbolism in early language acquisition. *Developmental science*, *21*(5), e12659.

[4] Shinohara, K., & Kawahara, S. (2010). A cross-linguistic study of sound symbolism: The images of size. In *Annual meeting of the berkeley linguistics society* (Vol. 36, No. 1, pp. 396-410).

[5] Sidhu, D. M., & Pexman, P. M. (2018). Five mechanisms of sound symbolic association. *Psychonomic bulletin & review*, *25*, 1619-1643.

[6] Adelman, J. S., Estes, Z., & Cossu, M. (2018). Emotional sound symbolism: Languages rapidly signal valence via phonemes. *Cognition*, *175*, 122-130.

[7] Winter, B., & Perlman, M. (2021). Size sound symbolism in the English lexicon. *Glossa: a journal of general linguistics*, *6*(1).

[8] Kilpatrick, A. J., Ćwiek, A., & Kawahara, S. (2023). Random forests, sound symbolism and Pokémon evolution. *PloS one*, *18*(1), e0279350.

[9] Li, Y., Zhao, Z., Klejch, O., Bell, P., & Lai, C. (2023). ASR and Emotional Speech: A Word-Level Investigation of the Mutual Impact of Speech and Emotion Recognition. *arXiv preprint arXiv:2305.16065*.

[10] G.G. Scott, A. Keitel, M. Becirspahic, et al. The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behav Res 51*, 1258–1270 (2019). https://doi.org/10.3758/s13428-018-1099-3.

[11] S.M. Mohammad, P.D. Turney, NRC Emotion Lexicon. *National Research Council Canada, 2*, 234 (2013).

[12] CMU Pronouncing Dictionary. (n.d.). Carnegie Mellon University. Retrieved June 16, 2023, from http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[13] R Core Team. *R: A language and environment for statistical computing*. (2023) [Computer software].

[14] T. Chen, T. He. *XGBoost: Extreme Gradient Boosting*. R package version 1.5.0.1. (2021) [Computer software].

[15] M. Kuhn. *caret: Classification and Regression Training*. R package version 6.0-88. (2023) [Computer software].

[16] S.A. Stouffer. *The American Soldier: Adjustment During Army Life* (Vol. 1). Princeton University Press. (1949).

[17] R.A. Fisher. *Statistical methods for research workers. Oliver and Boyd*. (1925).