Media-Analytics.org: A Resource to Research Language Usage by News Media Outlets

David Rozado [1*]

[1]ECL, Otago Polytechnic, 9054 Dunedin, New Zealand

Abstract. Most citizens in modern liberal democracies regularly consume news media content to inform themselves about current affairs. Thus, content analysis of news and opinion articles from popular media outlets can provide rich insight about the cultural milieu where such textual artifacts originated. Combining computational tools for content analysis with human led finesse can help an analyst exploit the capabilities of scalable computational methods while also leveraging human skills and expertise to guide the analysis. This work introduces an online tool, the http://media-analytics.org website, that empowers researchers by providing modern analytics tools to study language usage in textual content from news and opinion articles of major media outlets. Due to the diachronic nature of news articles, the Media-Analytics.org website allows the exploration of temporal dynamics in word frequency usage and strength of association between word pairs. It is the hope of the author that these tools can help other researchers gain insights about the temporal flux in language usage by major news media organizations.

1 INTRODUCTION

This work introduces and demonstrates the capabilities of an online research tool: The Media Analytics website (http://media-analytics.org), which allows an analyst to investigate language usage in diachronic news and opinion articles from popular news media outlets spanning up to 50 years of temporal coverage. The tool empowers an analyst with modern language analytics tools such as outlet-specific metrics about word frequency counts and strength of word pairs associations.

Human driven content analysis can generate detailed understanding about the themes contained in a corpus of natural language, but it is limited by humans' slow throughput when processing large volumes of text. Additionally, content analysis of subjective themes embedded in natural language constructs suffers from low intercoder reliability [1]. Computational content analysis can circumvent some of the limitations of human driven content analysis due to its reliability and scalability but it has its own set of limitations such as an inability to deal with the nuance of complex or subtle themes [2]. Combining computational tools with human led content analysis finesse can help an analyst exploit the capabilities of scalable computational methods while also leveraging human skills and expertise to guide the analysis [2]. The Media-Analytics.org website strives to facilitate that symbiotic relationship by providing a set of computational tools and metrics that empower a human analyst to test hypotheses about word usage in news media textual content.

* Corresponding author: david.rozado@op.ac.nz

2 METHODS

2.1 Sources of news articles

The metrics and models available at Media-Analytics.org were derived from a representative sample of news and opinion articles produced by popular news media outlets. News articles textual content are available in outlet-specific domains and Internet cache repositories such as the Internet Archive Wayback Machine, Google Cache and Common Crawl. Articles' headlines and main body texts can be located in HTML raw data using outlet-specific xpath expressions.

2.2 Frequency counts

News articles text was tokenized into unigrams (i.e. single words). No metrics are available for n-grams larger than 1. Several metrics have been computed for each word: The yearly absolute count of a word in an outlet represents the number of times the word occurs in headlines and body texts of all articles within a given year. The frequency of a single word in an outlet on any given year is estimated by dividing the count of all occurrences of the word in all articles of that year by the total count of all words in all articles of that year. The rank of a word in a year for a given outlet represents the frequency rank of a word in the entire vocabulary of that year.

2.3 Embedding models

Word embeddings (aka word vectors) are dense numerical representations of the syntactic and semantic roles of words in text corpora. The success of word embeddings in language modelling emerges from their ability to map the statistical cooccurrence of words and their contexts in the training corpus into positions in vector space that capture the semantic and syntactic roles with which the words are used in the corpus [3]. After training a word embedding model on a large volume of natural language, the resulting embedding space contains semantically meaningful spatial structure such that words that tend to cooccur in similar contexts are positioned in nearby regions of vector space.

A variety of algorithms to estimate optimal word vectors have been proposed over time. The Word2vec algorithm [2] consists of a swallow neural network trained to predict contextual words likely to appear in the vicinity of a target word. Two architectural choices for word2vec exist: The continuous bag of words (CBOW) version is trained to predict a target word from an aggregate of surrounding contextual words. The Skip gram (SG) version is trained to predict a single contextual word from a single target word. Other algorithmic techniques to estimate word embeddings have been proposed, such as Glove [4], but its performance is similar to word2vec. More recently, FastText [5] incorporates subword morphological information into the word embedding representations and outperforms word2vec in association and analogy tasks.
One of the main limitations of all these word embedding techniques has been their inability to distinguish different senses of a word (for instance financial *bank* from river *bank*). That is, word2vec, Glove and FastText, all bundle into one vector representations all the different senses with which a word might be used in a corpus. Very recently, contextual word embeddings techniques such as ELMo [6] or BERT [7] dynamically estimate a word vector representation based on the sentence in which it occurs thus capturing the sense with which a polysemous word is being used.

For the media analytics website, word2vec embedding models were trained on outlet-specific news articles at five-year time intervals within the 1970-2019 time range. To generate each word embedding model, the gensim [8] implementation of word2vec was used. The continuous bag of words (CBOW) architecture performed slightly better than the Skip-Gram architecture in commonly used validation metrics so it was used for subsequent analysis.

The ability of outlet-specific embedding models to capture semantic similarity, relatedness (i.e. association) as well as morphological, lexical, encyclopedic and lexicographic analogies [9] was measured, see Table 1. The performance of the embeddings derived from individual news outlets was roughly similar to several popular pre-trained embedding models trained on larger corpora such as a Twitter corpus or Google books. Performance of embeddings trained on news outlets content was only slightly worse than some famous popular pre-trained embedding models such as word2vec trained on Google News. This is due to said popular pre-trained embedding models being trained on corpora at least 2 orders of magnitude larger in size than the individual news outlets corpora used in this work. The FastText model trained on Common Crawl slightly outperformed all other embedding models probably due to it being trained on a very large training corpus and FastText ability to model morphological relationships at the subword level.

For training the outlet-specific word embedding models, the following parameters were used: vector dimensions=300, window size=10, negative sampling=10, down sampling frequent words = 0.0001, minimum frequency count for inclusion in vocabulary = 5, number of training iterations (epochs) through the corpus=5. The exponent used to shape the negative sampling distribution was 0.75.

Since the nature of the associations contained in a word embedding model is a function of the associations contained in the corpus on which the model was trained [10], [11], training on culturally distinct corpora along the dimensions of time or geographical location results in distinct latent associations that are consistent with the peculiarities of the corpus on which the embedding model was trained [2], [12]. Thus, a diachronic set of outlet-specific embedding models can be used as a proxy to trace the dynamics of word associations in the cultural and temporal context where the news articles were produced.

Table 1. Performance comparison between popular pre-trained embedding models (green) and word embedding models trained on outlet-specific news articles text (yellow) across commonly used validation metrics for embedding models.

	Popular pre-trained embedding models					Embedding models trained on news outlets articles				
Word embedding algorithm	word2vec (Skip-Gram)	word2vec (Skip-Gram)	Glove	Glove	fastext	word2vec (CBOW)	word2vec (CBOW)	word2vec (CBOW)	word2vec (CBOW)	word2vec (CBOW)
Vector dimensions	300	300	300	200	300	300	300	300	300	300
Training corpus name	Google News	Google Books N-grams 1990s	Wikipedia and Gigaword	Twitter	Common Crawl	New York Times (2015- 2019)	New York Times (1975- 1979)	Los Angeles Times (2015- 2019)	The Guardian (2015- 2019)	New York Post (2015- 2019)
Number of tokens in corpus	100B	NA	6B	27B	600B	255M	230M	719M	295M	119M
Model vocabulary size	3M	100K	400K	1.2M	2M	200K	190K	300K	217K	123K
WordSim-353	0.62	0.64	0.60	0.54	0.66	0.62	0.60	0.61	0.64	0.60
MEN simiarity dataset	0.68	0.68	0.74	0.61	0.71	0.71	0.71	0.69	0.74	0.66
SimLex-999	0.45	0.33	0.39	0.15	0.46	0.42	0.42	0.37	0.42	0.39
Google Semantic analogies	0.75	0.44	0.78	0.50	0.88	0.66	0.47	0.57	0.62	0.46
Google Syntactic analogies	0.74	0.39	0.67	0.60	0.84	0.61	0.57	0.53	0.65	0.52
BATS1 Inflectional Morphology analogies	0.68	0.36	0.60	0.51	0.85	0.51	0.45	0.44	0.51	0.44
BATS2 Derivational Morphology analogies	0.17	0.05	0.09	0.08	0.32	0.07	0.08	0.06	0.09	0.06
BATS3 Encyclopedic Semantics analogies	0.21	0.15	0.25	0.18	0.30	0.16	0.12	0.14	0.17	0.12
BATS4 Lexicographic Semantics analogies	0.06	0.08	0.07	0.07	0.10	0.04	0.05	0.04	0.04	0.03
AVERAGE	**0.48**	**0.35**	**0.47**	**0.36**	**0.57**	**0.42**	**0.39**	**0.38**	**0.43**	**0.36**

3 MEDIA-ANALYTICS.ORG FEATURES AND APPLICATIONS

3.1 Frequency counts

The usage of word frequencies to study cultural phenomena is validated next by showing how frequency metrics properly track historical events as well as shifting societal trends [13]. **Fig. 1** shows min-max normalized word frequencies of several words in New York Times articles and opinion pieces for the time range 1970-2019. Normalized min-max frequency metrics are obtained using the formula $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$ with x representing the absolute frequency of a word in a year, x_{min} representing the minimum frequency of the word in the time series and x_{max} representing the maximum frequency of the word in the time series. After applying this normalization operation, the output is guaranteed to lie in the 0-1 range.

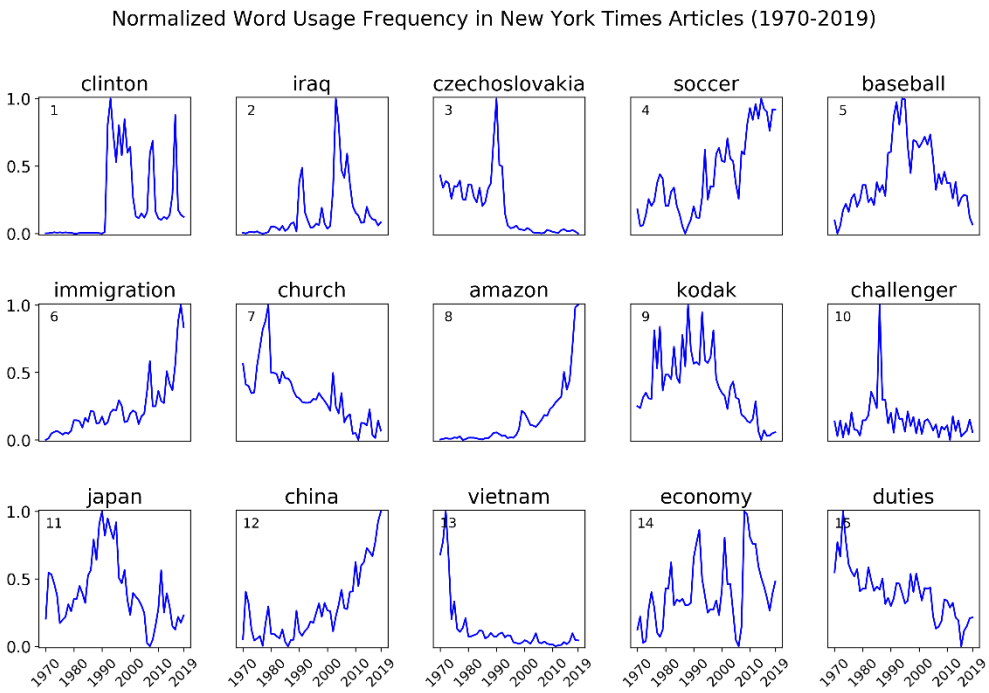

Fig. 1. Normalized min-max frequency counts of several sample words in New York Times articles

Plotting absolute frequency counts across several outlets allows an analyst to compare the prevalence of certain topics across different media outlets. **Fig. 2** shows absolute frequency counts of several sample words across four different news outlets. Clear differences and commonalities in the absolute prevalence of different words across outlets are apparent in the figure.

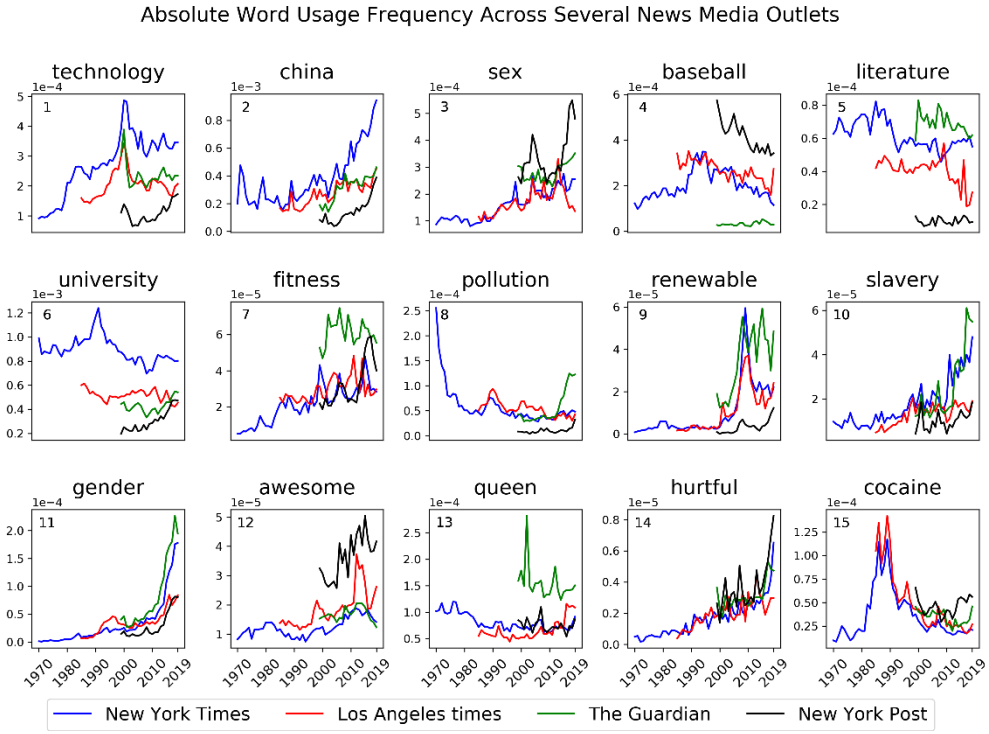

Fig. 2. Absolute yearly frequencies of several words across four different news media outlets

3.2 Correlation of frequency counts across outlets

In order to allow flexible and innovative usage of the metrics and models available at the media analytics website, the media-analytics.org app allows the metrics provided to be downloaded in CSV format for subsequent analysis with additional software tools. One example usage of this raw data could calculate correlations of frequency counts across outlets. Correlation of frequency counts for a word across news outlets can serve to visualize the degree of synchronicity with which different outlets choose to cover a certain topic.

Fig. *3* shows that some news outlets are highly coupled in how frequently they talk about the sample word *baseball*, while others seem to follow a more independent trajectory.

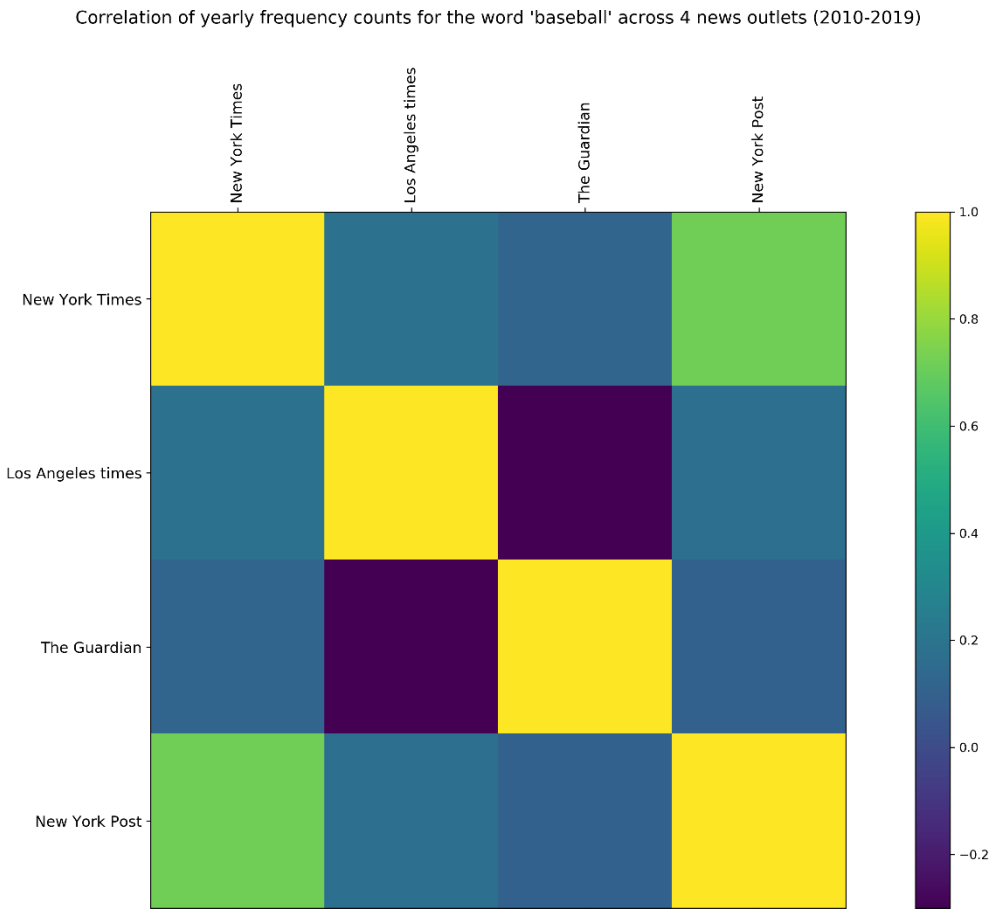

Fig. 3. Correlation matrix of frequency counts for the word *baseball* across 4 news outlets

4.6 Word associations using word embeddings

Diachronic word embedding models trained on outlet specific news articles distil changing latent associations between word pairs in outlet content. This metric can be used to visualize how word associations flow through time in news media content.

Fig. *4* displays such associations in New York Times content for a selected sample of word pairs. Several societal and historical trends are clearly apparent in the Figure.

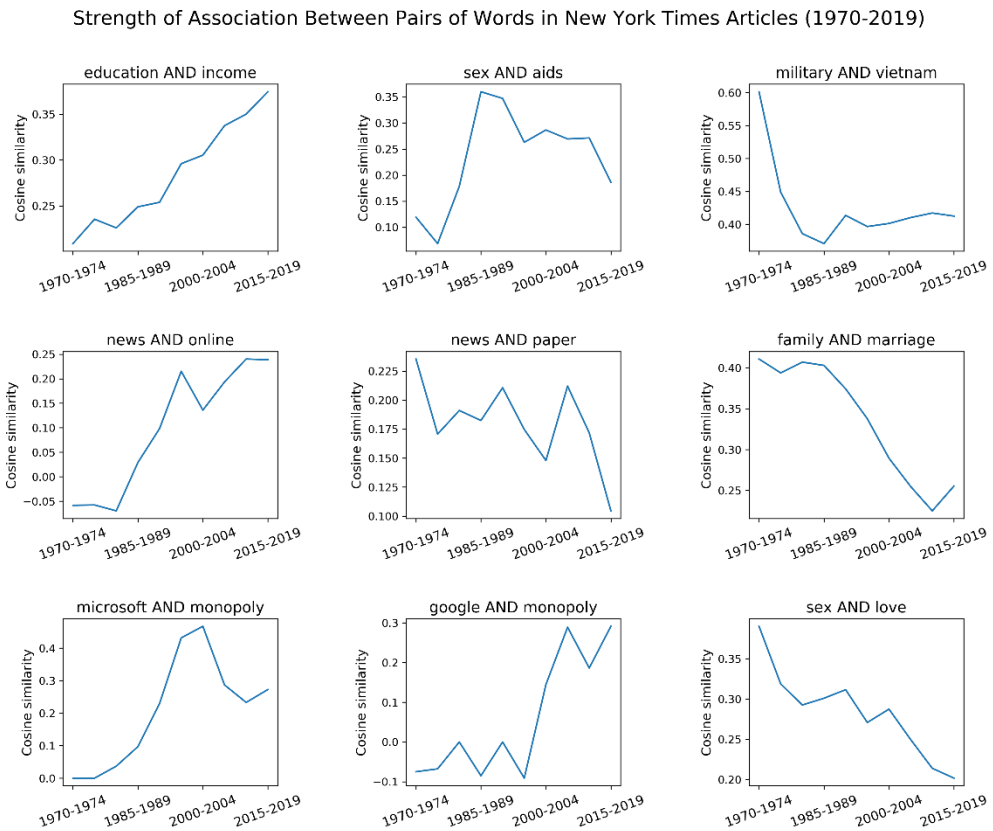

Fig. 4. Strength of association between word pairs in a diachronic corpus of NYT articles.

References

1. K. A. Neuendorf, SAGE Publications, 1st edition, (2001)
2. A. C. Kozlowski, M. Taddy, J. A. Evans, *Am Sociol Rev*. **84**, 905–949 (2019)
3. T. Mikolov, W. Yih, G. Zweig, *CNACACL*, pp. 746–751 (2013)
4. J. Pennington, R. Socher, C. Manning, *EMNLP*, pp. 1532–1543 (2014)
5. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching, TACL **5**, 135–146 (2017)
6. M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, *arXiv:1802.05365 [cs]* (2018)
7. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *arXiv:1810.04805 [cs]* (2018)
8. R. Řehůřek, P. Sojka, *LREC* pp. 45–50 (2010)
9. A. Gladkova, A. Drozd, S. Matsuoka, *NAACL*, pp. 8–15 (2016)
10. D. Rozado, *PLOS ONE*. **15**, e0231189 (2020)
11. D. Rozado, *Soc*. **56**, 256–266 (2019)
12. N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, *PNAS*. **115**, E3635–E3644 (2018)
13. D. Rozado, *Acad. Quest.* **33**, 89–100 (2020)