nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-019-0112-6

A universal information theoretic approach to the identification of stopwords

Martin Gerlach[1,5*], Hanyu Shi[1,5] and Luís A. Nunes Amaral[1,2,3,4*]

One of the most widely used approaches in natural language processing and information retrieval is the so-called bag-of-words model. A common component of such methods is the removal of uninformative words, commonly referred to as stopwords. Currently, most practitioners use manually curated stopword lists. This approach is problematic because it cannot be readily generalized across knowledge domains or languages. As a result of the difficulty in rigorously defining stopwords, there have been few systematic studies on the effect of stopword removal on algorithm performance, which is reflected in the ongoing debate on whether to keep or remove stopwords. Here we address this challenge by formulating an information theoretic framework that automatically identifies uninformative words in a corpus. We show that our framework not only outperforms other stopword heuristics, but also allows for a substantial reduction of document size in applications of topic modelling. Our findings can be readily generalized to other bag-of-words-type approaches beyond language such as in the statistical analysis of transcriptomics, audio or image corpora.

The use of methods from natural language processing[1] has become an indispensable tool in applications of data science pervading nearly every scientific discipline[2,3]. The main challenge is how to extract meaningful information from large and diverse datasets—most of which are comprised of unstructured texts. One of the most common approaches to represent textual data is the so-called bag-of-words model, in which one ignores the order of words within a given document. To improve the signal-to-noise ratio or decrease the amount of data, this is often accompanied by data filtering as part of the data pre-processing steps[4]. In practice, such activities can take up to 80% of the research effort[5]. However, we still lack fundamental insights into how these procedures affect the performance of specific algorithms[6].

For concreteness, we consider topic modelling[7], a paradigmatic unsupervised approach for automatic organization of collections of documents[8]. One contentious pre-processing step in topic modelling is the removal of semantically uninformative words such as 'the'. The most common approach, which goes back more than 50 years, is to curate a "dictionary of insignificant words"[9], commonly referred to as a stopword list[10]. While some stopword lists can appear to practitioners as standard due to being the default choice in popular applications (such as Mallet[11]), there is no consensus among experts on which words should be excluded[12].

Indeed, the use of a 'standard' stopword list is problematic because it ignores the domain-knowledge specificity of stopwords[13] and because it is language-specific[14]. The limitation of static lists has motivated the development of other heuristic approaches based on factors such as the number of occurrences (most and least frequent words), document frequency, and term frequency and inverse document frequency (TFIDF)[15], and other, often ill-specified, procedures. The state of uncertainty in the field is illustrated by the fact that the seminal paper on latent Dirichlet allocation "removed a standard list of 50 stop words ... [and] ... words that occurred only once"[16], but other works by the same author subsequently removed "standard stop words and those that appear too frequently or too rarely"[17] or "all words not in a pruned vocabulary of 4,253 words"[18], or chose "1,539 terms that occurred in more than five documents"[19] or a "5,000-term vocabulary according to tfidf"[20]. Even when using the same method, such as TFIDF, different authors use different thresholds; for example, ref. [21] removes words "that have tfidf greater than 0.8". The inconsistency in filtering approaches poses severe challenges to the comparison of results across different studies, rendering it nearly impossible to obtain a coherent picture on the state of the field. This is exacerbated by the fact that the removal of stopwords in topic modelling and text-based unsupervised learning more generally is not well understood[22], leading to a sterile debate on the usefulness of such approaches.

Model

Inspired by the formulation of Montemurro and Zanette[23], we define a metric that quantifies how uninformative a word is in a corpus by using the framework of information theory.

Conditional entropy. Consider a corpus $\mathcal{C}$ with D documents in total. We denote by $n(w,d)$ the number of occurrences (tokens) of a word w in document d such that the number of tokens in document d is $n(d)=\sum_w n(w,d)$, and $n(w)=\sum_d n(w,d)$ is the frequency of word w. It follows that the total number of tokens in the entire corpus is $N=\sum_{w,d} n(w,d)$. For each word w, we consider its distribution over documents as:

$$p(d|w)=\frac{p(w,d)}{p(w)}=\frac{n(w,d)}{n(w)} \quad (1)$$

where $p(w)=n(w)/N$ is the relative frequency of a word. Using the Shannon entropy, we can quantify how 'uneven' this distribution is:

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA. [2]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, USA. [3]Department of Physics and Astronomy, Northwestern University, Evanston, IL, USA. [4]Department of Medicine, Northwestern University, Evanston, IL, USA. [5]These authors contributed equally: Martin Gerlach, Hanyu Shi. *e-mail: mgerlach@wikimedia.org; amaral@northwestern.edu

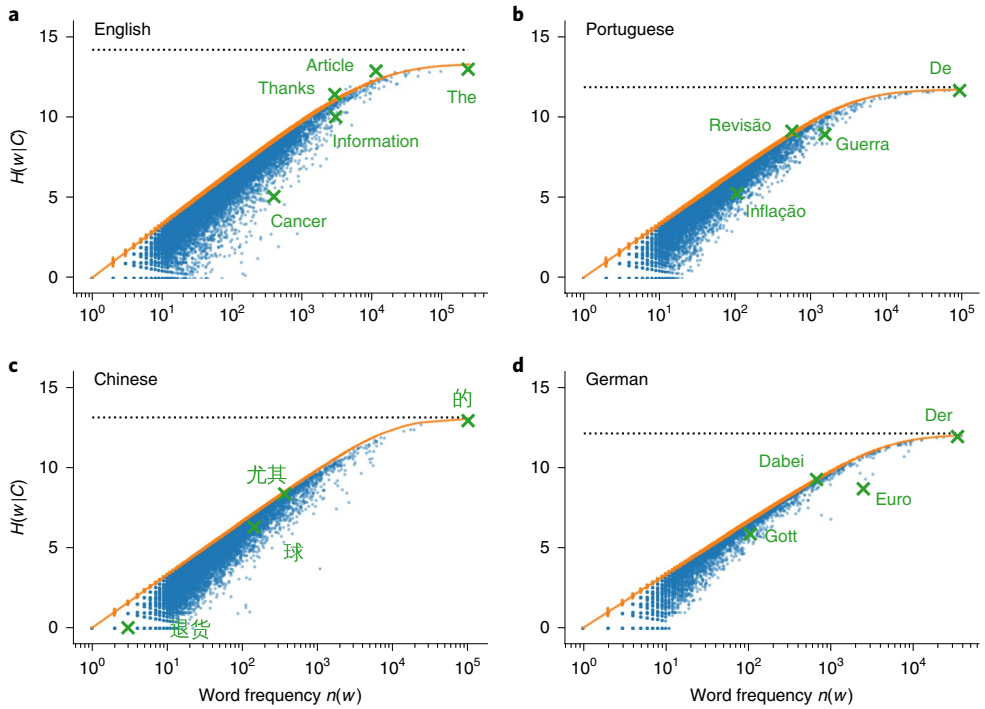

Fig. 1 | Using entropy as a universal measure to quantify the information content of a word. The conditional entropy $H(w|\mathcal{C})$ as a function of the frequency $n(w)$ for each word (blue dots) and the expected entropy $\langle \widetilde{H}(w|\mathcal{C})\rangle$ from a random null model averaged over 1,000 realizations (orange line) for different languages (see Supplementary Methods A for details on the datasets). The error bars represent five s.d. **a**, English (20 newsgroups). **b**, Portuguese. **c**, Chinese. **d**, German. The maximum entropy $H_{\max} = \log D$ (for example, ≈14.2 for English) is shown as a dotted line. Individual words are shown as examples (crosses).

$$H(w|\mathcal{C}) = -\sum_d p(d|w)\,\log p(d|w) \tag{2}$$

This conditional entropy can be interpreted as a dispersion measure[24] quantifying in bits the amount of uncertainty a randomly drawn token of word w provides about which document d it occurs in. Thus, a more informative word will have a lower conditional entropy. For example, for an extremely topical word that occurs in only one document d^*, we have $p(d|w) = \delta_{d,d^*}$ and we obtain $H(w|\mathcal{C}) = 0 = H_{\min}$. Here δ is the Kronecker delta: $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. In contrast, an idealized stopword would be evenly distributed, yielding $H(w|\mathcal{C}) = \sum_d 1/D\,\log 1/D = \log D = H_{\max}$.

Random null model. In reality, the last term provides only an upper bound for the entropy of idealized stopwords due to finite-size effects. In particular, Zipf's law for word frequencies[25] tells us that most words occur with a very low frequency and we will, thus, be expected to undersample $p(d|w)$.

To correct for undersampling, we construct a null model to estimate the expected entropy of randomly distributed words, $\widetilde{H}(w|\mathcal{C})$. Specifically, by shuffling all tokens across documents, we obtain a random distribution of words across documents $\tilde{n}(w,d)$ while preserving the marginal counts $n(w)$ and $n(d)$. Note that an alternative random null model in which each word is used with fixed relative frequency $p(w)$ yields indistinguishable results (see Supplementary Notes A and Supplementary Fig. 1). As expected, we find that $\widetilde{H}(w|\mathcal{C})$ depends strongly on $n(w)$ (Fig. 1a); its functional form can be roughly approximated by

$$\widetilde{H}(w|\mathcal{C}) \propto \log\left(1 - \mathrm{e}^{-n(w)/D}\right) \tag{3}$$

(see Supplementary Notes B and Supplementary Fig. 2). An implication of this result is that using a fixed threshold entropy value to determine stopword lists will inevitably eliminate informative words and include uninformative ones.

For the most common words, such as 'the' in English corpora, the observed entropy approaches the null model value ($H \lesssim \widetilde{H}$), but still remains slightly smaller than $\log D$. Moreover, there are many words with medium or low frequency $n(w)$ whose occurrence is indistinguishable from chance. In fact, some words have $H > \widetilde{H}$ (that is, they are more equally distributed than we would expect from chance), which can be attributed to a highly regular usage across documents (for example, 'thanks'). Significantly, words that have $H \ll \widetilde{H}$, such as 'cancer' in an English corpus, can be statistically distinguished from words that are used randomly across all documents.

Importantly, we observe very similar results for corpora from other languages (Fig. 1b–d). High-frequency function words do display only small deviations from the expected entropy (for example, 'de' in Portuguese, 的 in Chinese or 'der' in German). However, many words with medium or low frequency appear in the corpus in a manner indistinguishable from the prediction of the null model. Yet, we also observe a substantial set of words that have smaller than expected entropies such as 'inflação' in Portuguese, 球 in Chinese or 'gott' in German.

Information content. Prompted by these findings, we define the information content of a word as the difference between the observed and the expected values of the conditional entropy

$$I(w|\mathcal{C}) \equiv \langle \widetilde{H}(w|\mathcal{C})\rangle - H(w|\mathcal{C}) \tag{4}$$

where $\langle \widetilde{H}(w|\mathcal{C})\rangle$ is the average over different realizations of the null model. A word with $I(w|\mathcal{C}) \approx 0$ is statistically indistinguishable from a word that is used at random. Thus, low values of $I(w|\mathcal{C})$ can be used to identify stopwords.

Remarkably, for the studied corpora, the vast majority of words turn out to be uninformative, which can be rationalized by the fact

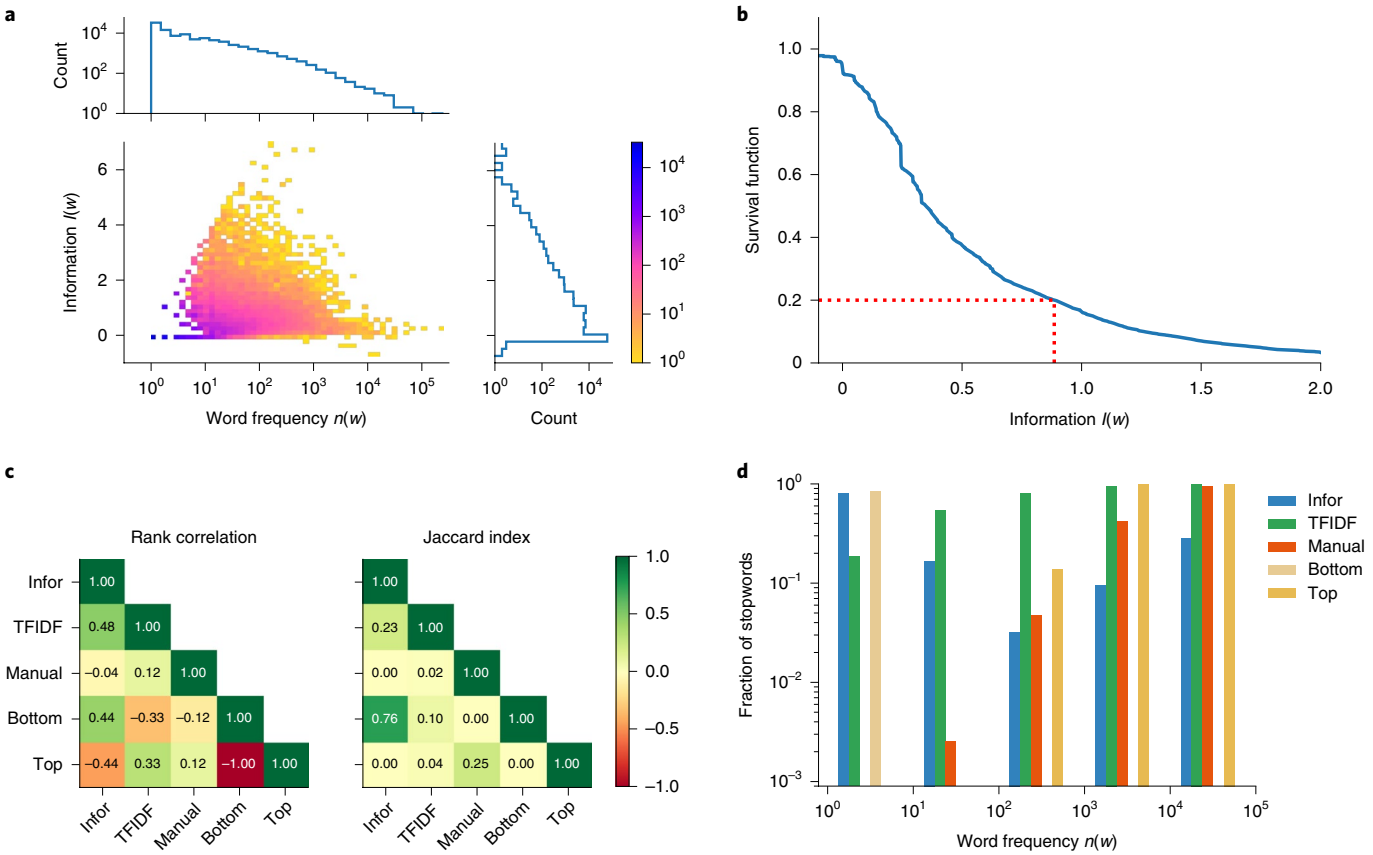

Fig. 2 | Identification of stopwords by thresholding of information content. a, The number of words with a given value of information content $I(w)$ and frequency $n(w)$ for the English corpus (20 newsgroup corpus) in Fig. 1a. The histograms (top and right) show the marginal counts. **b**, The survival function (that is, the number of tokens remaining after removing all tokens of words with an information content smaller than $I(w)$). The dotted red line indicates the information content for which 80% of tokens are removed. **c**, A comparison of stopword lists using different methods. Left, Spearman's rank correlation coefficient between scores $S_i(w)$ and $S_j(w)$ for methods i and j. Right, the Jaccard index among the resulting stopword lists from methods i and j using fixed thresholds. **d**, The fraction of words identified as stopwords conditioned on their frequency. In **c** and **d** we use thresholds for Infor (0.1), TFIDF (9) and Bottom (5), whereas for Top we filter the 1,000 words with the largest frequency (see Supplementary Methods B for details).

that most words occur very rarely (Fig. 2a). The joint distribution over $I(w)$ and $n(w)$ reveals a more intricate pattern (Fig. 2a and Supplementary Fig. 3). For the 77% of the words that occur fewer than 10 times, we have a typical information content close to 0. Only for words occurring more than ten times do we find a substantial fraction of words with non-zero information content. Yet, even in this case, the peak of the distribution of $I(w|\mathcal{C})$ is typically located near 0. This indicates that uninformative words can be found across the full spectrum of word frequencies. Interestingly, words with a large number of occurrences usually deemed uninformative can show small deviations from random usage $I(w) \approx 0.2$ (meaning that they are in fact informative) or negative values (indicating a more regular usage than would be generated by chance).

We deem a deviation from the null expectation statistically significant if it exceeds a threshold value (Supplementary Fig. 4). In general, this threshold depends on $n(w)$. However, we found that it was smaller than 0.1 bits across all word frequencies (one-sided P value = 0.05). While this selection imposes that the information content is significantly larger than 0, it does not guarantee the magnitude of the information content.

Therefore, our approach to defining stopwords is any word with an absolute information content smaller than $I^* > 0.1$. Interestingly, scanning of I^* does not reveal an obvious optimum as the number of remaining tokens decays exponentially (Fig. 2b and Supplementary Fig. 5). Thus, the choice of I^* can be based on the desired reduction in the size of the data—a moderate threshold $I^* = 1$ removes more than 80% of the data in both the number of words and tokens.

Experiments

Next, we benchmark our principled approach against currently popular stopword identification approaches (Supplementary Methods B).

Overlap with other heuristics. We start by quantifying the overlap between different stopword lists using the correlation score as well as the Jaccard index (Supplementary Methods C). While both calculations appear to show a significant overlap between our approach and TFIDF (Fig. 2c), a subtler analysis reveals large differences (Fig. 2d). Indeed, taking into account the uneven distribution of words (Zipf's law), our approach is the only one that does show a non-monotonic U-shape removing mostly words with high and low frequency.

Topic model inference. While the pragmatic reason to remove stopwords is to reduce the computational cost of the topic model inference (Supplementary Fig. 6), the most exciting potential of stopword removal in topic modelling is an improvement in the quality of the inferred topics. A major challenge in the evaluation of topic models in this context is that common metrics such as perplexity and coherence are ill-suited to assess the effect of removing stopwords as they require unchanged data (Supplementary Notes C). In fact, it has

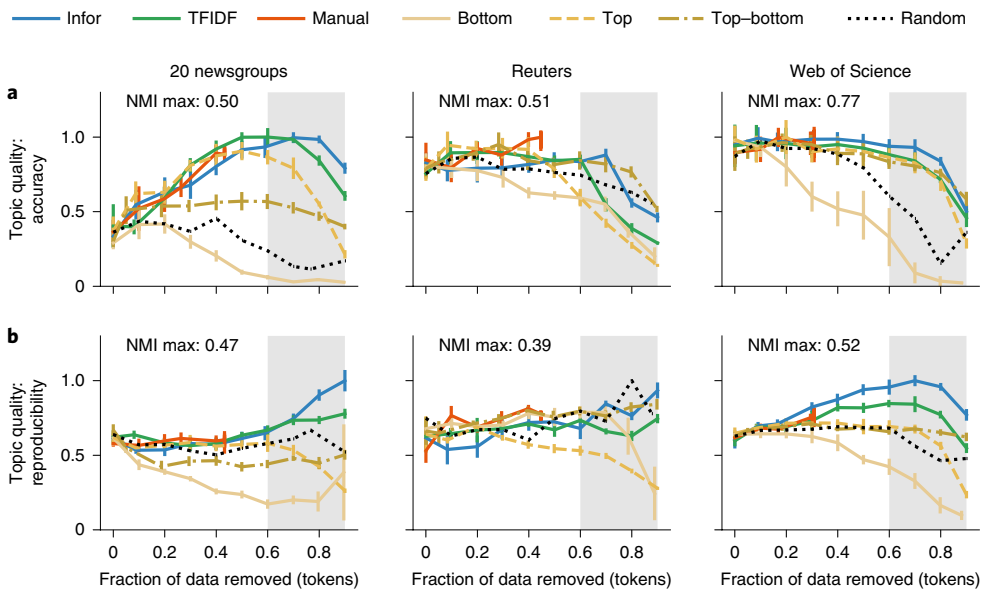

Fig. 3 | Removal of information theoretic stopwords makes the topic model more accurate and stable. The performance of a topic model algorithm as a function of the fraction of removed stopwords using different stopword lists for three English corpora from different knowledge domains (Supplementary Methods A). **a**, Accuracy (that is, the overlap between the inferred topic distribution and the metadata document labels as measured via NMI; see Supplementary Methods D). **b**, Reproducibility (that is, the overlap of the topic labels of words between inference results from different 'runs' of the same topic model algorithm and the same corpus as measured via NMI (also known as token clustering[27])). For ease of visualization, both measures are reported as ratios with respect to the maximum value of NMI observed in each plot (we report this characteristic scale as NMI_{max}). The curves show the average and ±2 s.d. over 10 different inference runs with the HDP topic model (Supplementary Methods E); see Supplementary Figs. 7–9, which show similar results for other topic model algorithms. The shaded areas indicate the regions in which the information theoretic approach yields the best results across all corpora and topic models and, importantly, corpus size reductions are greatest.

been observed empirically that "traditional topic quality metrics are not robust to stopwords"[26] because they are implicitly biased by, for example, the size of the vocabulary[27] (Supplementary Figs. 13–15).

As a first proxy for the quality of the inferred topics, we assess their accuracy by quantifying how much the inferred topic distributions correlate with category labels from the document metadata. Specifically, we calculate the normalized mutual information (NMI) between the documents' topic (that is, the topic with highest propensity) and their category labels. For the case of synthetic data—the only case in which we have access to a well-defined ground-truth topic structure—this measure has been shown to correlate strongly with measures quantifying the overlap between planted and inferred topics[28]. While for real corpora, metadata can and should not be treated as a ground truth in absolute terms[29], an increase or decrease in the NMI can highlight a positive or negative effect on the quality of topic model inference when removing stopwords, avoiding the biases observed for traditional measures. As a second proxy for the quality of the inferred topics, we quantify the reproducibility of the inferred topic distributions across different realizations of the optimization algorithm given the same data. In practice, the inferred topics can vary due to different initial conditions (in the case of variational Bayes) or stochasticity in the inference algorithm (in the case of Gibbs sampling), leading to different local maxima in the likelihood landscape. While this effect is not well understood, it has been shown that the different local maxima can correspond to substantially different solutions[30]. Therefore, a high overlap between different solutions would mitigate such issues, and corroborate the robustness of the inferred topics.

Our information theoretic approach leads to substantial improvements, both in accuracy and in reproducibility, across different corpora (Fig. 3), while at the same time reducing the amount of data by as much as 80%. Remarkably, manually curated and TFIDF's stopword lists perform almost as well as our information theoretic approach when removing only a small fraction of the data, indicating that these historically grown lists constitute a good heuristic. However, the information theoretic approach typically yields the maximum performance compared to alternatives when removing a very large fraction of data (between 60% and 80%), a range that is inaccessible to manual approaches and in which performance substantially deteriorates for TFIDF. Interestingly, the effect of stopword heuristics can vary dramatically for different topic model algorithms (Supplementary Figs. 7–9). While individual heuristics work well in a specific scenario, our information theoretic approach is the only approach that consistently displays high performance across different evaluation metrics, corpora and topic models. For example, the 'Top–bottom' heuristic (removing high- and low-frequency words) highly outperforms our information theoretic approach for some corpora in combination with the LDAVB topic model. However, even then, LDAVB yields weaker overall performance than the other topic model algorithms (HDP or LDAGS).

Next, we investigate the potential of applying our approach to corpora from different languages. While manual stopword lists for English are readily available, it is not only time-consuming but also challenging to compile such lists for an arbitrary language[14]. Our approach offers a scalable alternative as it automatically identifies words that do not contain any informative content in a statistical sense.

Considering annotated corpora from three different languages (Portuguese, Chinese and German), we largely reproduce the results obtained for English (Fig. 4 and Supplementary Figs. 10–12 for other topic model algorithms). For these three corpora, removal of stopwords according to our proposed measure leads to topics that are both more accurate and more reproducible. Similarly to the results for English corpora, the information theoretic approach yields the maximum performance when removing large fractions of the data. In particular, for the Chinese and German corpora, it leads to a large

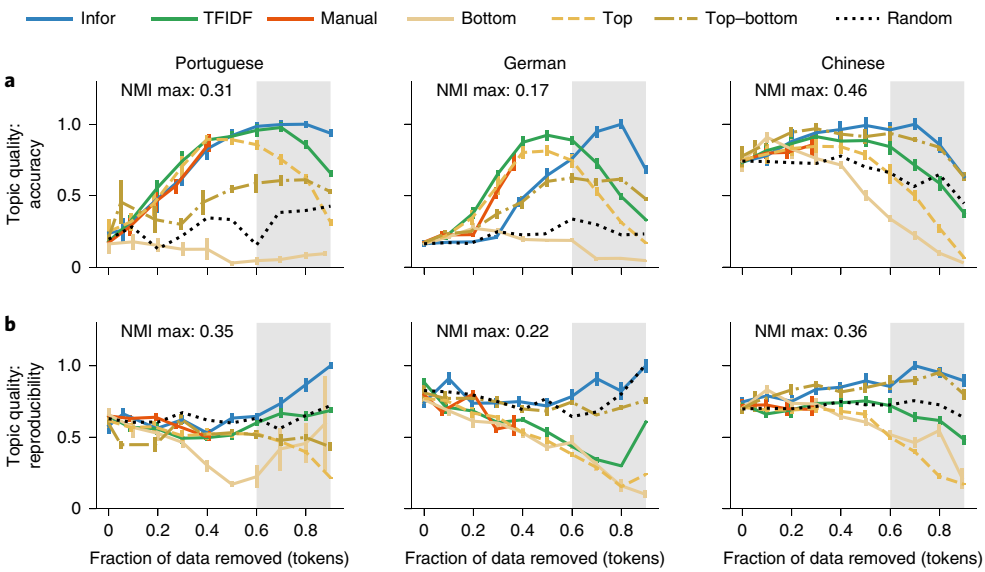

Fig. 4 | Universal improvement of topic model inference for different language corpora. The performance of a topic model algorithm as a function of the fraction of removed stopwords according to different stopword lists for three corpora from different languages (Supplementary Methods A). **a**, Accuracy. **b**, Reproducibility. For ease of visualization, both measures are reported as ratios with respect to the maximum value of NMI observed in each plot (we report this characteristic scale as $\mathrm{NMI}_{\mathrm{max}}$). The curves show the average and ±2 s.d. over 10 different inference runs with the HDP topic model (Supplementary Methods E); see Supplementary Figs. 10–12, which show similar results for other topic model algorithms and additional stopword lists. The shaded areas indicate the regions in which the information theoretic approach yields the best results across all corpora and topic models and, importantly, corpus size reductions are greatest.

improvement in the overlap between topic distributions obtained from different runs of the topic model.

Generalizablity to other bag-of-words models. The information theoretic approach to the identification of stopwords introduced here is not only advantageous in applications of topic modelling, but it also has high potential for bag-of-words approaches in general. To support this claim, we consider here two case studies.

First, we consider the problem of document classification in information retrieval[31]. Specifically, we follow the approach in ref. [32] and investigate the performance of stopword removal for the supervised classification of document labels using support vector machines with word counts $n(w,d)$ as document features. In Fig. 5, we show the prediction accuracy (the fraction of correctly classified documents; normalized by the maximum accuracy across all different stopwords) in a held-out test set with tenfold cross-validation. The removal of stopwords does not increase accuracy, but allows for a substantial reduction in the size of the corpus with little decrease in accuracy for small and intermediate fractions of removed data (<60%). Even without any filtering, accuracy is close to 1, showing that supervised classification of category labels is a much easier problem than unsupervised inference such as topic modelling. The differences between different stopword lists become amplified if the fraction of removed data is large (>60%). Surprisingly, among the traditional stopword lists, performance can vary strongly across corpora. Most importantly however, the information theoretic stopword list is the only approach that yields the maximum performance across corpora from any domain or language.

Second, we show that our method can be generalized to cases beyond language by applying it to data from single-cell RNA-sequencing (scRNA-seq)[33], where one counts the number of times a gene (equivalent to a word) is expressed in individual cells (equivalent to a document), or more precisely the number of gene-specific RNA molecules. In analogy to the bag-of-words model for texts, computational analysis of data from scRNA-seq experiments aims to automatically identify different cell types[34]. Different pre-processing heuristics are employed to filter stopwords (that is, so-called housekeeping genes that are required for basic cell functioning and are consistently expressed across all cells[35]). In Fig. 6a we quantify the informativeness of 17,467 genes based on their expression profiles across 713 individual cells taken from a single human donor (Supplementary Methods A). Inspection of the conditional entropy (H) and its expected value from a random null model ($\widetilde{H}$) reveals the same patterns we observed for textual data (Fig. 1). The expression counts of most genes across cells are indistinguishable from chance. In particular, for the most common genes such as *MALAT1*, we find that $H \approx \widetilde{H}$ and thus they are deemed uninformative. In contrast, genes such as *GNLY*[36] or *PTGDS*[37] with an intermediate overall number of counts exhibit $H \ll \widetilde{H}$ and thus constitute the most informative genes. Curated annotations on the biological role of genes are consistent with our classification: whereas uninformative genes are vital for basic cell functioning (for example, scaffolds), examples of informative genes are associated with more specific contexts in particular cell types such as T cells (Fig. 6b).

These examples offer a new view on the effect of stopword removal according to our information theoretic approach. While a gene might be considered uninformative for cell type identification, that gene is probably important for the survival of the cell. Similarly, words that are uninformative for the topic identification are nonetheless vital for the readability of a text (Fig. 6c).

Discussion

In contrast to other heuristics, our proposed method substantially reduces the total amount of data as well as the size of the vocabulary and it can be applied without any fine-tuning in corpora originating from different knowledge domains or languages. While our analysis is confined to bag-of-words approaches, the formulation grounded on information theory allows for straightforward extensions that take into account additional structural features such as sentences, paragraphs or the context windows used in, for example, word2vec approaches[38]. Given the wide use of 'topic model' approaches in biology[39] or image

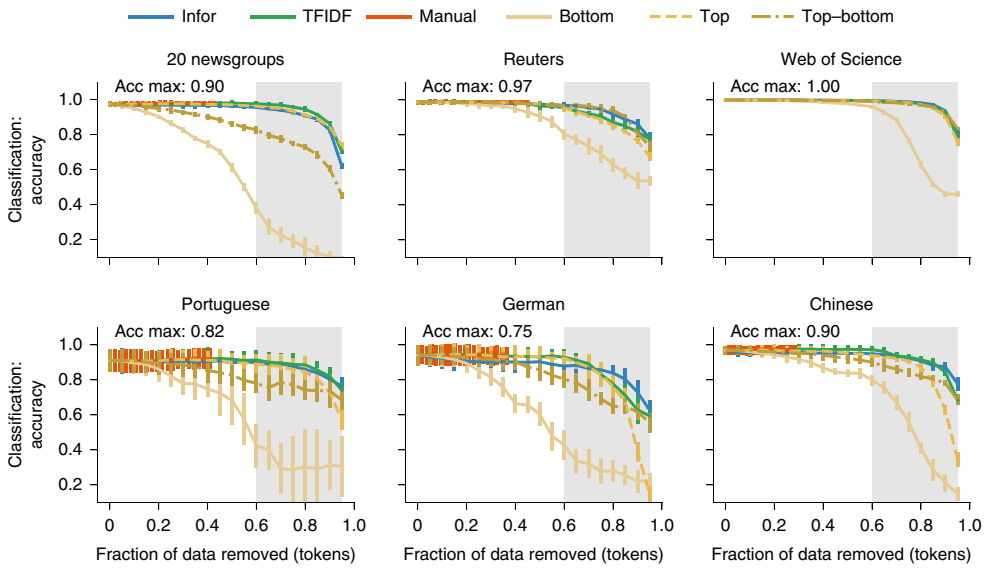

Fig. 5 | Robustness of supervised classification accuracy with respect to removal of information theoretic stopwords. The accuracy in supervised document classification (using support vector machines in a bag-of-words model with word counts as document features) as a function of the fraction of removed stopwords, using different stopword lists for corpora from different knowledge domains (top row) and languages (bottom row). The curves show the average and ±2 s.d. from tenfold cross-validation. The shaded areas indicate the regions in which the information theoretic approach yields the best results across all corpora.

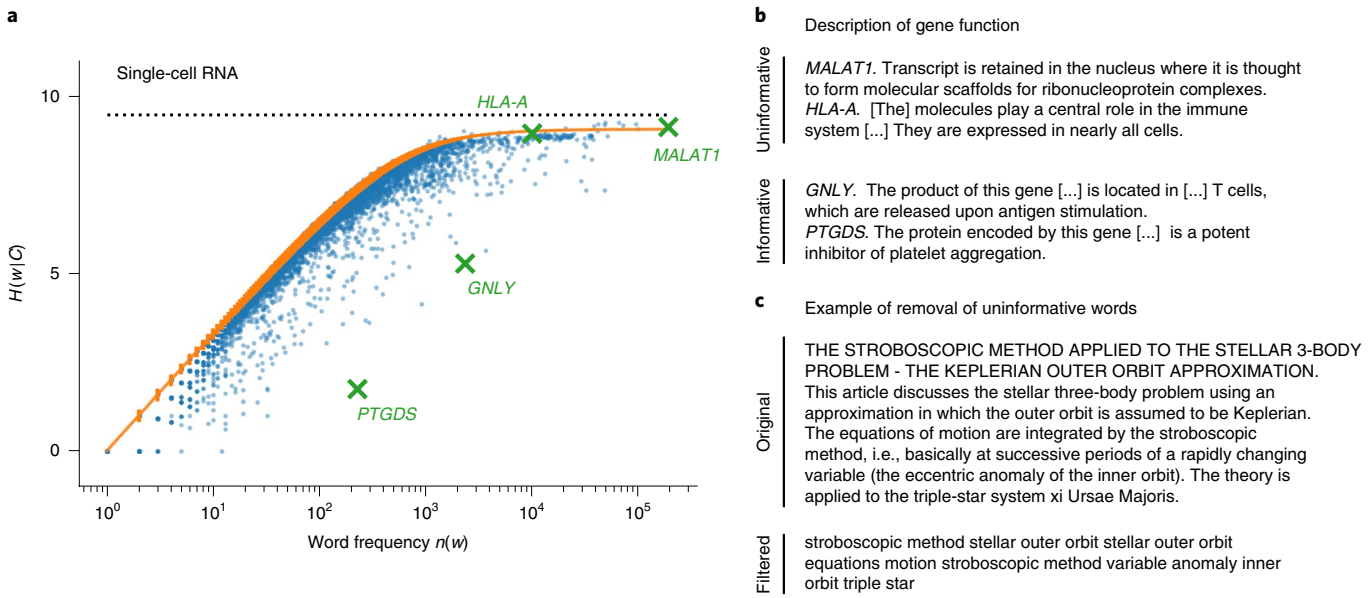

Fig. 6 | Application to data from scRNA-seq reveals 'stopgenes'. a, The conditional entropy $H(w|\mathcal{C})$ as a function of the frequency $n(w)$ for each gene (blue dots) and the expected entropy $\langle\tilde{H}(w|\mathcal{C})\rangle$ from a random null model averaged over 1,000 realizations (orange line) for a dataset from scRNA-seq that measures the number of times a gene is expressed in a cell (see Supplementary Methods A for details on the datasets). The error bar represents two s.d. **b**, A description of the function of four example genes as stated in https://www.ncbi.nlm.nih.gov/gene, showing the consistency with our classification as informative and uninformative. **c**, The effect of stopword removal for an example text (the abstract of a scientific paper from the Web of Science) when removing 70% of word tokens with the information theoretic approach (Infor). Although the filtered text is not readable, its content is reduced to keywords that highlight the differences to all other texts contained in the corpus.

analysis[40] facing similar issues in pre-processing, our method could be applicable beyond the analysis of texts and it can be seen as a principled approach to the common problem of thresholding[41].

Data availability

The text data are available in the public repository https://github.com/amarallab/stopwords.

Code availability

The code for this Article, along with an accompanying computational environment, is available in the public repository https://github.com/amarallab/stopwords and is executable online as a Code Ocean capsule. Code for the calculation of the information theoretic measure I and for the experiments with topic models can be found at https://doi.org/10.24433/CO.6204149.v1[42].

Received: 7 March 2019; Accepted: 9 October 2019;
Published online: 02 December 2019

References

1. Manning, C. D. & Schütze, H. *Foundations of Statistical Natural Language Processing* (MIT Press, 1999).
2. Evans, J. A. & Aceves, P. Machine translation: mining text for social theory. *Ann. Rev. Sociol.* **42**, 21–50 (2016).
3. Rebholz-Schuhmann, D., Oellrich, A. & Hoehndorf, R. Text-mining solutions for biomedical research: enabling integrative biology. *Nat. Rev. Genet.* **13**, 829–839 (2012).
4. García, S., Luengo, J. & Herrera, F. *Data Preprocessing in Data Mining* (Springer, 2014).
5. Dasu, T. & Johnson, T. *Exploratory Data Mining and Data Cleaning* (John Wiley & Sons, 2003).
6. Schoenfeld, B., Giraud-Carrier, C., Poggemann, M., Christensen, J. & Seppi, K. Preprocessor selection for machine learning pipelines. Preprint at http://arXiv.org/abs/1810.09942 (2018).
7. Blei, D. M. Probabilistic topic models. *Commun. ACM* **55**, 77–84 (2012).
8. Boyd-Graber, J., Hu, Y. & Mimno, D. Applications of topic models. *Found. Trends Inf. Retr.* **11**, 143–296 (2017).
9. Luhn, H. P. The automatic creation of literature abstracts. *IBM J. Res. Dev.* **2**, 159–165 (1958).
10. Rasmussen, E. in *Encyclopedia of Database Systems* (eds Liu, L. & Özsu, M. T.) (2009).
11. McCallum, A. K. Mallet: a machine learning for language toolkit. http://mallet.cs.umass.edu (2002).
12. Nothman, J., Qin, H. & Yurchak, R. Stop word lists in free open-source software packages. In *Proc. Workshop for NLP Open Source Software (NLP-OSS)* (eds Park, E. L. et al.) 7–12 (Association for Computational Linguistics, 2018).
13. Lo, R. T.-W., He, B. & Ounis, I. Automatically building a stopword list for an information retrieval system. *J. Digit. Inf. Manag.* **5**, 17–24 (2005).
14. Zou, F., Wang, F. L., Deng, X., Han, S. & Wang, L. S. Automatic construction of Chinese stop word list. In *Proc. 5th WSEAS International Conference on Applied Computer Science (ACOS'06)* (Huang, W. et al.) 1009–1014 (World Scientific and Engineering Academy and Society, 2006).
15. Salton, G. & Yang, C. S. On the specification of term values in automatic indexing. *J. Doc.* **29**, 351–372 (1973).
16. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
17. Wang, C., Paisley, J. & Blei, D. M. Online variational inference for the hierarchical Dirichelet process. In *Proc. 14th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research* Vol. 15, 752–760 (AISTAT, 2011).
18. Hoffman, M. D., Blei, D. M. & Bach, F. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)* (eds Lafferty, J. D. et al.) 1–9 (Neural Information Processing Systems Foundation, 2010).
19. Blei, D. M., Griffiths, T. L. & Jordan, M. I. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM* **57**, 1–30 (2010).
20. Blei, D. M. & Mcauliffe, J. D. Supervised topic models. In *Advances in Neural Information Processing Systems* (eds Platt J. C. et al.) vol. 20, 121–128 (NIPS 2007).
21. Achakulvisut, T., Acuna, D. E., Ruangrong, T. & Kording, K. Science concierge: A fast content-based recommendation system for scientific publications. *PLoS ONE* **11**, e0158423 (2016).
22. Schofield, A., Magnusson, M. & Mimno, D. Pulling out the stops: rethinking stopword removal for topic models. In *Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics* (eds Lapata, M. et al.) Vol. 2, 432–436 (Association for Computational Linguistics, 2017).
23. Montemurro, M. A. & Zanette, D. H. Towards the quantification of the semantic information encoded in written language. *Adv. Complex Syst.* **13**, 135–153 (2010).
24. Gries, S. T. Dispersions and adjusted frequencies in corpora. *Int. J. Corpus Linguist.* **13**, 403–437 (2008).
25. Zipf, G. K. *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, 1949).
26. Fan, A., Doshi-Velez, F. & Miratrix, L. Prior matters: simple and general methods for evaluating and improving topic quality in topic modeling. Preprint at http://arXiv.org/abs/1701.03227 (2017).
27. Schofield, A. & Mimno, D. Comparing apples to apple: the effects of stemmers on topic models. *Trans. Assoc. Comput. Linguist.* **4**, 287–300 (2016).
28. Shi, H., Gerlach, M., Diersen, I., Downey, D. & Amaral, L. A new evaluation framework for topic modeling algorithms based on synthetic corpora. In *Proc. Machine Learning Research* Vol. 89 (eds. Chaudhuri, K. & Sugiyama, M.) 816–826 (PMLR, 2019).
29. Peel, L., Larremore, D. B. & Clauset, A. The ground truth about metadata and community detection in networks. *Sci. Adv.* **3**, e1602548 (2017).
30. Lancichinetti, A. et al. High-reproducibility and high-accuracy method for automated topic classification. *Phys. Rev. X* **5**, 011007 (2015).
31. Aggarwal, C. C. & Zhai, C. in *Mining Text Data* (eds. Aggarwal, C. C. & Zhai, C.) 77–128 (Springer, 2012).
32. Uysal, A. K. & Gunal, S. The impact of preprocessing on text classification. *Inf. Process. Manag.* **50**, 104–112 (2014).
33. Skinnider, M. A., Squair, J. W. & Foster, L. J. Evaluating measures of association for single-cell transcriptomics. *Nat. Methods* **16**, 381–386 (2019).
34. Bravo González-Blas, C. et al. Cistopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
35. Alberts, B. et al. *Molecular Biology of the Cell* Sixth International Student Edition (W. W. Norton & Co., 2014).
36. Zheng, C. et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* **169**, 1342–1356.e16 (2017).
37. Solé-Boldo, L. et al. Single-cell transcriptomes of the aging human skin reveal loss of fibroblast priming. Preprint at *bioRxiv* https://doi.org/10.1101/633131 (2019).
38. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26* (eds Burges, C. J. C. et al.) 3111–3119 (Curran Associates, 2013).
39. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
40. Broderick, T., Mackey, L., Paisley, J. & Jordan, M. I. Combinatorial clustering and the beta negative binomial process. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 290–306 (2015).
41. Yan, X., Jeub, L. G. S., Flammini, A., Radicchi, F. & Fortunato, S. Weight thresholding on complex networks. *Phys. Rev. E* **98**, 042304 (2018).
42. Gerlach, M., Shi, H. & Amaral, L. A. N. Stopwords-filtering. *Code Ocean* https://doi.org/10.24433/CO.6204149.v1 (2019).

Acknowledgements

L.A.N.A. acknowledges a John and Leslie McQuown Gift to NICO and support from the Department of Defense Army Research Office (grant number W911NF-14-1-0259). M.G. thanks T. Stoeger and Z. Ren for insightful discussion on scRNA-seq.

Author contributions

M.G. and L.A.N.A. conceptualized the study. M.G. and H.S. obtained all data and conducted all analysis. M.G. and L.A.N.A. wrote the first draft. M.G., H.S. and L.A.N.A. edited and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-019-0112-6.

Correspondence and requests for materials should be addressed to M.G. or L.A.N.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.