Software Impacts 12 (2022) 100270

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

Fumeus: A family of Python tools for text mining with smoke terms

David M. Goldberg [a,*], Richard J. Gruss [b], Alan S. Abrahams [c]

[a] San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, United States of America
[b] Radford University, P.O. Box 6954, Radford, VA 24142, United States of America
[c] Virginia Tech, 880 West Campus Drive, Blacksburg, VA 24061, United States of America

ARTICLE INFO

Keywords:
Text mining
Natural language processing
Information retrieval
Machine learning
Decision support
Product safety

ABSTRACT

Synthesizing meaningful insights from voluminous textual datasets is complex and challenging. The task is especially difficult for sparse target classes. Recent works have proposed "smoke terms," or machine-learned words and phrases prevalent in a target class. Smoke terms may be utilized to rank or sort text, or they may serve as features in follow-on machine learning models. This paper introduces Fumeus, a family of Python-based smoke term analysis tools. We provide functionality to generate new smoke terms and to use existing smoke term dictionaries to rank or sort datasets. These analyses have numerous academic, regulatory, and industry applications.

Code metadata

Current code version	*v1*
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2022-19
Permanent link to reproducible capsule	https://codeocean.com/capsule/5287429/tree/v1
Legal code license	*MIT License*
Code versioning system used	*None*
Software code languages, tools and services used	*Python*
Compilation requirements, operating environments, and dependencies	*Natural Language Toolkit (NLTK), Beautiful Soup (BS4), Chardet*
If available, link to developer documentation/manual	https://github.com/fumeus/fumeus
Support email for questions	fumeus@protonmail.com

1. Introduction

The growth of online media and electronic word of mouth has enabled the collection of rich troves of textual data. However, despite the immense potential of textual datasets to create value, their volumes pose difficulties for decision-makers seeking to synthesize insights. The information systems literature terms this problem "information overload" [1,2], which Hemp [3] describes as a blurring of the line between valuable information and distracting information. Deriving digestible insights from these datasets remains a challenging problem and the subject of ongoing research efforts.

Several recent works have proposed the use of "smoke terms" to address this challenge. Smoke terms are words and phrases that are especially prevalent in a target class of interest and infrequent otherwise. This methodology arose from research efforts to mine product defect-related information from online consumer reviews [4,5]. Defect-related discussions are immensely valuable for industry and regulatory decision-makers, but they are sparse enough that manual review of the text is impractical. Industry-specific smoke terms such as "airbag" served as excellent predictors of the target class "safety concerns" in the automotive industry [4,5], and recent works have extended this methodology to other sectors, such as baby products [6], toys [7], joint and muscle aids [8], food [9], and appliances [10,11]. Recent efforts have also extended this methodology beyond product-related complaints to areas such as hospital services [12] and financial services [13,14].

Comparative analyses have found that smoke terms often outperform out-of-the-box approaches such as sentiment analysis when seeking to detect sparse target classes [6,10,15–17]. Comparative analyses for several applications have found that smoke terms were narrowly outperformed by the performance of deep learning word

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: dgoldberg@sdsu.edu (D.M. Goldberg), rgruss@radford.edu (R.J. Gruss), abra@vt.edu (A.S. Abrahams).

https://doi.org/10.1016/j.simpa.2022.100270
Received 18 February 2022; Received in revised form 22 February 2022; Accepted 9 March 2022

embedding models [13,14]. However, a substantial advantage of smoke terms over these models is their interpretability. Rather than a "black box" approach whose inner workings are opaque to decision-makers, smoke terms can be emphasized in textual summaries to indicate the model's process [13,14]. "Interpretable artificial intelligence" has been an emerging area in the literature in recent years [18], with Sachan et al. [19] finding that models lacking in interpretability face obstacles to adoption.

In this paper, we present a new family of Python tools for smoke term analysis called Fumeus (Latin for "smokey"). Fumeus enables rich smoke term analyses for numerous academic, government, and industry text analysis applications.

2. Software description

Fumeus is available in two forms. First, we make our Python source code available so that research and practitioners may directly utilize it for their analyses. These Python scripts may be integrated into existing workflows, built upon for new applications, or adapted for further experiments. Second, we also provide a web graphical user interface (see the Fumeus GitHub page for details), which allows users to perform the same analyses without the need to engage with the Python source code directly. Uploaded files are deleted after use, and neither settings nor results are otherwise cached, maintaining the confidentiality of analyses. Fumeus offers two core functions: smoke term generation and smoke term scoring.

2.1. Smoke term generation

Fumeus's first core function involves generating smoke terms for any arbitrary textual dataset. Three lengths of smoke terms are supported: unigrams (words), bigrams (two-word phrases), and trigrams (three-word phrases). In addition, four information retrieval metrics are supported to derive smoke terms: "Correlation Coefficient" ("CC"), "Robertson's Selection Value" ("RSV"), "Relevance Correlation Value" ("RSV"), and "Document and Relevance Correlation" ("DRC") [20]. Fumeus returns a list of the highest-scoring smoke terms via the chosen metric as well as the corresponding scores for the smoke terms returned. With all metrics, the highest-scoring smoke terms have the highest relative prevalence (according to that metric) in the target class versus the contrast (non-target) class. See Goldberg and Abrahams [10] for a detailed discussion of this methodology. Fig. 1 shows a schematic of the smoke term generation process.

Results are sorted by the chosen information retrieval metric in descending order and are available in Comma-Separated Value (CSV) or JavaScript Object Notation (JSON) formats. Fig. 2 shows an example of smoke term generation with Fumeus using an example dataset and output.

2.2. Smoke term scoring

Fumeus's second core function involves computing smoke term-based scores for unseen records (textual narratives). Both a textual dataset and a weighted smoke term dictionary are required to perform this analysis. The smoke term dictionary may be a direct output of Fumeus's smoke term generation, or alternative smoke term dictionaries, such as sentiment dictionaries, may be utilized instead. In computing smoke term-based scores, each occurrence of a smoke term increments the score by its weight, and records are then sorted in descending order. As Goldberg and Abrahams [10] discuss, in a subsequent step, this score may be normalized by word count to avoid a bias toward longer records. However, as several studies have found [10,14], this normalization does not necessarily improve performance, so Fumeus does not perform this normalization. Fumeus returns the unnormalized scores for each record as well as a list of all smoke terms found in each record. Results are sorted by unnormalized scores in descending order and are available in Comma-Separated Value (CSV) or JavaScript Object Notation (JSON) formats. Fig. 3 shows the process for smoke term scoring with an example dataset, dictionary, and output.

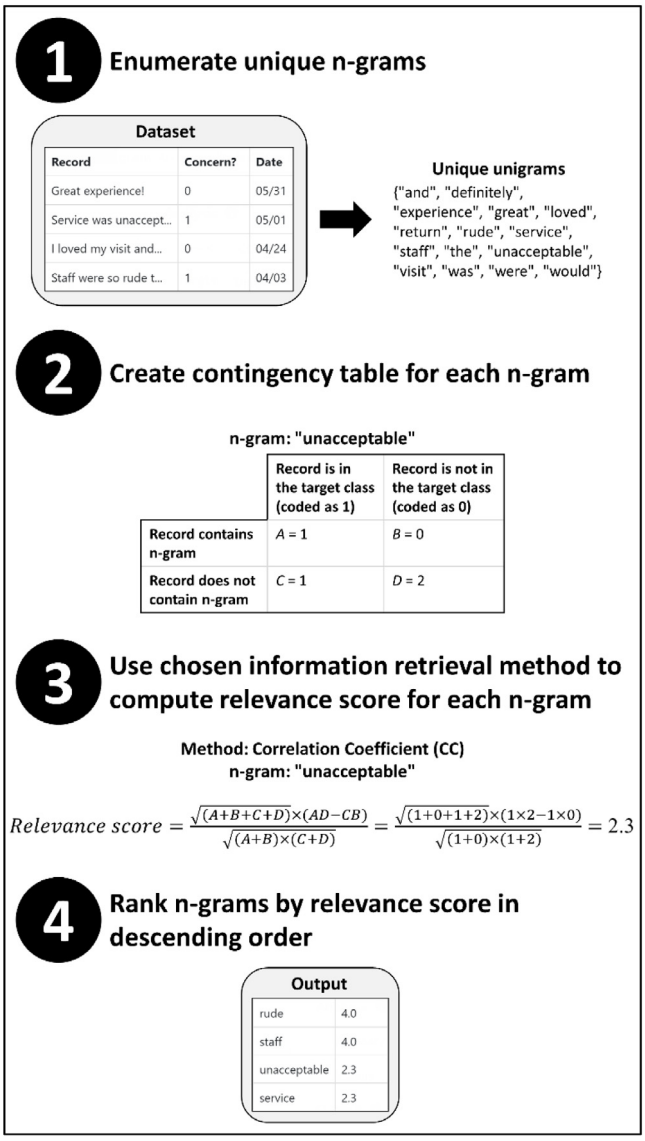

Fig. 1. Smoke term generation process.

3. Impacts

3.1. Uses of smoke terms for text mining

The generated smoke terms and corresponding weights have several possible uses for text analyses. Several studies use the smoke terms directly to rank or sort records [4–6,8,11,15] (as mentioned above, a follow-on normalization step may be appropriate [10]). Records can then be ranked by this score in descending order, where high-scoring records are most likely to be of the target class. In this application, the smoke term model does not serve as a classifier, but rather a prioritization tool for decision-makers. A decision-maker may manually assess the highest-ranking records, which are most likely to be relevant, and continue reading in descending order until they assess that they have exhausted the valuable records. Metrics such as normalized discounted cumulative gain [21] have been utilized in past works to assess the quality of these rankings [9]. Previous works have also shown that smoke terms may be stylized (bolded, highlighted, etc.) in each record to communicate the model's inner workings to the decision-maker [13,14]. Fig. 4 shows an example application in which smoke terms were developed to identify mentions of safety hazards

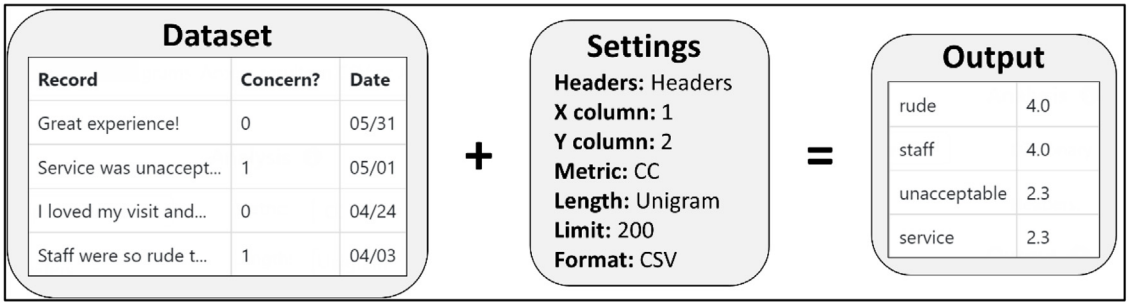

Fig. 2. Fumeus smoke term generation.

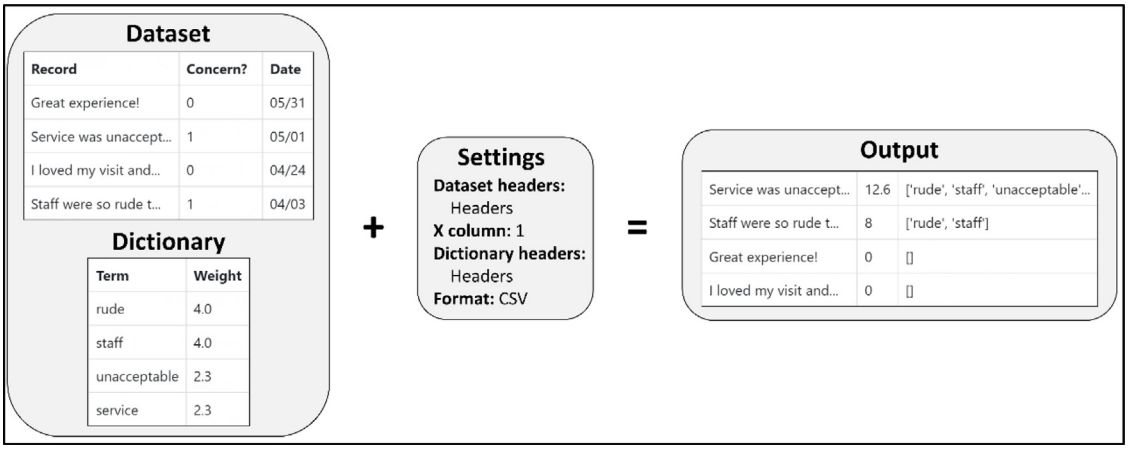

Fig. 3. Fumeus smoke term scoring.

Confidence of Safety Hazard	Date	Rating	Review
99	6/14/2016	1	BREAD CAUGHT ON FIRE ARE YOU KIDDING ME. My bread literally caught on fire this morning. ACTUAL FIRE. The setting wasn't even on its highest, it was at a 3. Imagine if I had walked away, my entire kitchen would've burned down. This is extremely dangerous and unacceptable. I don't know how you make your toasters but they're not supposed to catch on fire. It's ridiculous.
97	10/15/2014	1	Door jambed closed and the bread caught fire. had ... Door jambed closed and the bread caught fire. had to break the door to put the fire out. No it's not air tight, so pushing the door closed didn't put the fire out. Dangerous product. Buy at own risk..

Fig. 4. Example smoke term-based interface.

in countertop appliance reviews [10]. Smoke terms are emphasized in red colour, bolded, and underlined. Rather than exclusively showing the decision-maker the confidence that each review refers to a safety hazard (for example, with a deep learning word embeddings model), the emphasized smoke terms improve interpretability by demonstrating the model's rationale.

A further use of smoke terms is as features in follow-on machine learning models. As large corpora contain many unique n-grams, dimensionality reduction is a substantial concern for text-based machine learning models [22]. The highest-weighted smoke terms are likely to be strong predictors, so they may be used to reduce the dimensionality of text and provide suitable vectors for machine learning models. Brahma et al. [13] and Goldberg et al. [14] each used smoke terms for this purpose, and both found that follow-on machine learning models approached the accuracy of deep learning word embedding models. Additional studies have used heuristic methods to fine-tune candidate smoke terms to maximize precision [9,10,16].

3.2. Application areas and future work

Historically, a major application of smoke terms has been in identifying mentions of product defects in online discussions, such as consumer reviews or other discussion fora [4–6,9–11,15]. In this context, mentions of product defects are relatively uncommon, and smoke terms provide a means of ranking records to aid in review by a decision-maker. A subset of this research focuses on a more specific target class of safety defects, mentions of which are similarly high-valued for decision-makers but very sparse in online media [7,10,23]. Other applications have included hospital services [12] and financial services [13,14].

Fumeus has future applications in a variety of domains. For instance, Fumeus could be utilized to perform predictive analyses using medical records. Smoke terms within physicians' notes could be generated to predict adverse health outcomes, thus constructing an early warning system for future complications. Alternatively, Fumeus also has applications to mining terror chatter, where the target class may be very sparse in a large dataset. Smoke terms could dramatically reduce the dimensionality of the data, then enabling a decision-maker to consider a shortlist of records for closer analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P.J. Denning, ACM president's letter: Electronic junk, Commun. ACM 25 (3) (1982) 163–165.

[2] S.R. Hiltz, M. Turoff, Structuring computer-mediated communication systems to avoid information overload, Commun. ACM 28 (7) (1985) 680–689.

[3] P. Hemp, Death by information overload, Harv. Bus. Rev. 87 (9) (2009) 83–89.

[4] A.S. Abrahams, J. Jiao, W. Fan, G.A. Wang, Z. Zhang, What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings, Decis. Support Syst. 55 (4) (2013) 871–882.

[5] A.S. Abrahams, J. Jiao, G.A. Wang, W. Fan, Vehicle defect discovery from social media, Decis. Support Syst. 54 (1) (2012) 87–97.

[6] V. Mummalaneni, R. Gruss, D.M. Goldberg, J.P. Ehsani, A.S. Abrahams, Social media analytics for quality surveillance and safety hazard detection in baby cribs, Saf. Sci. 104 (2018) 260–268.

[7] M. Winkler, A.S. Abrahams, R. Gruss, J.P. Ehsani, Toy safety surveillance from online reviews, Decis. Support Syst. 90 (2016) 23–32.

[8] D.Z. Adams, R. Gruss, A.S. Abrahams, Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews, Int. J. Med. Inf. 100 (2017) 108–120.

[9] D.M. Goldberg, S. Khan, N. Zaman, R.J. Gruss, A.S. Abrahams, Text mining approaches for postmarket food safety surveillance using online media, Risk Anal. (2020).

[10] D.M. Goldberg, A.S. Abrahams, A tabu search heuristic for smoke term curation in safety defect discovery, Decis. Support Syst. 105 (2018) 52–65.

[11] D. Law, R. Gruss, A.S. Abrahams, Automated defect discovery for dishwasher appliances from online consumer reviews, Expert Syst. Appl. 67 (2017) 84–94.

[12] N. Zaman, D.M. Goldberg, A.S. Abrahams, R.A. Essig, Facebook hospital reviews: automated service quality detection and relationships with patient satisfaction, Decis. Sci. 52 (6) (2021) 1403–1431.

[13] A. Brahma, D.M. Goldberg, N. Zaman, M. Aloiso, Automated mortgage origination delay detection from textual conversations, Decis. Support Syst. 140 (2021) 113433.

[14] D.M. Goldberg, N. Zaman, A. Brahma, M. Aloiso, Are mortgage loan closing delay risks predictable? A predictive analysis using text mining on discussion threads, J. Assoc. Inf. Sci. Technol. (2021).

[15] A.S. Abrahams, W. Fan, G.A. Wang, Z. Zhang, J. Jiao, An integrated text analytic framework for product defect discovery, Prod. Oper. Manage. 24 (6) (2015) 975–990.

[16] D.M. Goldberg, A.S. Abrahams, Sourcing product innovation intelligence from online reviews, Decis. Support Syst. (2022).

[17] N. Zaman, D.M. Goldberg, R.J. Gruss, A.S. Abrahams, S. Srisawas, P. Ractham, M.M. Şeref, Cross-category defect discovery from online reviews: Supplementing sentiment with category-specific semantics, Inf. Syst. Front. (2021) 1–21.

[18] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215.

[19] S. Sachan, J.-B. Yang, D.-L. Xu, D.E. Benavides, Y. Li, An explainable AI decision-support-system to automate loan underwriting, Expert Syst. Appl. 144 (2020) 113100.

[20] W. Fan, M.D. Gordon, P. Pathak, Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison, Decis. Support Syst. 40 (2) (2005) 213–233.

[21] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Trans. Inf. Syst. 20 (4) (2002) 422–446.

[22] H. Kim, P. Howland, H. Park, N. Christianini, Dimension reduction in text classification with support vector machines, J. Mach. Learn. Res. 6 (1) (2005).

[23] L. Nasri, M. Baghersad, R. Gruss, N.S.W. Marucchi, A.S. Abrahams, J.P. Ehsani, An investigation into online videos as a source of safety hazard reports, J. Saf. Res. 65 (2018) 89–99.