Towards Automated Detection of Hype in Biomedical Research

Bojan Batalo[1*] Erica K. Shimomoto[1‡*] Neil Millar[2]
[1]National Institute of Advanced Industrial Science and Technology [2]University of Tsukuba
{bojan.batalo,kidoshimomoto.e}@aist.go.jp
{millar.neil}@u.tsukuba.ac.jp

Abstract

The use of promotional language ('hype') in biomedical research is increasing. Examples include adjectives such as *groundbreaking*, *unparalleled*, *novel* and *innovative*. Such language can undermine objective evaluation of evidence, impede development of research and erode trust in science. In this pilot study, we show that (1) formalizing annotation guidelines may help humans reliably annotate such adjectives as 'hype' or 'not hype', and (2) that using an annotated dataset following the guidelines to train machine learning models yields promising results for automatic detection of promotional language.

1 Introduction

The language of biomedical research is becoming increasingly promotional – a phenomenon referred to as 'hype' [1]. For example, investigators promote the significance and novelty of their research using exaggerated terms (e.g. *revolutionary*). They describe research problems in dramatic terms (*daunting*). They amplify the scale and rigor of their methods (*extensive*, *robust*) and the utility of the results (*actionable*, *impactful*).Increasing use of hype has been demonstrated in biomedical funding applications [2] and journal publications [3], while comparable trends are evident in other research fields [4]. Figure 1 shows examples of 'hype' in sentences.

Hype in science is a cause for concern. As the former editor-in-chief of JAMA Network journals points out, words such as *ground-breaking*, *transformative*, or *unprecedented* are rarely justified, and may undermine objective assessment, impeding the development of further studies, policies, clinical practice, and knowledge translation [5]. Moreover, promotional and confident language can bias readers' evaluation of research, [6, 7], and public trust in science is eroded when promotional language creates unrealistic expectations or misrepresents findings [8].

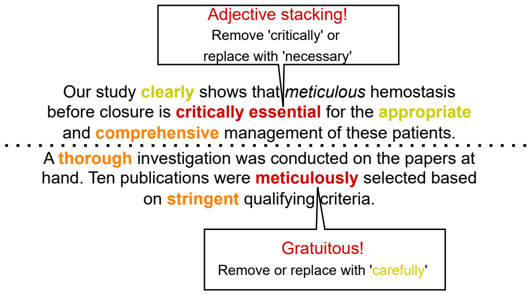

Adjective stacking!
Remove 'critically' or replace with 'necessary'
Our study clearly shows that meticulous hemostasis before closure is critically essential for the appropriate and comprehensive management of these patients.
A thorough investigation was conducted on the papers at hand. Ten publications were meticulously selected based on stringent qualifying criteria.
Gratuitous!
Remove or replace with 'carefully'

Figure 1: Example sentences containing heavily promotional adjectives. Stacking adjectives to increase the urgency or gratuitous amplification of research rigor are some examples of 'hype'.

To combat salesmanship in science, different approaches are needed, and among these, technological solutions (e.g., systems to detect, assess, and provide feedback on promotional language) may be one means to foster objectivity and accountability. However, whether a given word or phrase is promotional depends on the context. For instance, adjectives like *essential* and *meticulous* can promote significance or rigor, but they may also occur in a neutral context or technical phrase (e.g., essential fatty acid, meticulous hemostasis). At the same time, words with similar meanings can vary in promotional intensity (e.g., *new* vs. *novel* vs. *innovative* vs. *groundbreaking* etc.).

Previously, we defined the concept of hype as 'hyperbolic and/or subjective language that authors use to glamorize, promote, embellish and/or exaggerate aspects of their research'. We have created raw corpora of scientific texts [1, 2, 9], used these to identify a lexicon of 140

[*] Authors contributed equally to this work.

adjectives that can carry promotional meaning in biomedical texts [2, 10]. In that work, adjectives were deemed potentially hype if over 30% of occurrences were, after manually examining them, promotional.

Although our lexicon of terms can help identify candidates, discussions about whether a specific term constitutes hyper remain problematic. For annotators, determining whether a word is used with the intention to promote involves subjective judgment based on context and interpretation. Moreover, our overarching definition of hype has proved inadequate for distinguishing ambiguous cases, forcing annotators to rely on intuition and group discussion.

In this work, we propose some formal annotation guidelines that can be used to determine whether an adjective is used in a promotional manner based on its semantics, function, and context. Furthermore, we manually annotate a dataset of 550 sentences containing potentially promotional adjectives using these guidelines and discuss annotation disagreements.

Finally, we formulate hype classification as a text classification task, and test some classical natural language processing methods such as Multinomial Naive Bayes, Multivariate Bernoulli, Support Vector Machines, via bag-of-words approach and using word embeddings.

2 Dataset annotation

The starting point of our work is the corpus of raw texts compiled by Millar et al. [2], which comprises 901,717 abstracts of successfully funded grant applications submitted to and approved by the National Institutes of Health (NIH) in the United States. The NIH corpus has 335 million words, out of which are 36.4 million adjectives. In previous studies, Millar et al. [2] have identified 140 adjectives that they deemed to be ‘potentially hype’ according to a simple, broad guideline defined as: *“If the adjective has **positive value judgment**, and can be **removed or replaced without loss in meaning**, it is potentially hype”*.

The 140 identified ‘potentially hype’ adjectives have been divided into eight groups, based on the aspect of research they are promoting: **importance**, **novelty**, **rigour**, **scale**, **utility**, **quality**, **attitude**, **problem**. As the starting point, we select the **novelty** group of adjectives, which emphasize the novelty and innovation of proposed research, an aspect of research reinforced by the academic peer-review system and the competitive funding process. This adjective group comprises 11 members: *creative*, *emerging*, *first*, *groundbreaking*, *innovative*, *latest*, *novel*, *revolutionary*, *unique*, *unparalleled*, *unprecedented*.

To compile our dataset, we use the CQPweb [11] to search for novelty adjectives through the recent NIH corpus abstracts, covering years from 2016 to 2020. This search yielded a total of 161,469 occurrences, covering 84,299 abstracts. Due to time and resource constraints, we limit ourselves to a smaller corpus, which can be annotated and manually examined by the three authors in a reasonable time frame. We randomly choose 50 samples per adjective, resulting in 550 samples, covering 545 abstracts.

2.1 Annotation guidelines

One of the authors, Neil Millar, a linguist, designed the initial annotation guidelines based on his experience and expertise in linguistics and hype research. The guidelines comprise several steps that might require high proficiency in the English language but are designed to be easy to follow and can be applied sequentially. We assume looking at an adjective within the context of a sentence; with this starting point, the annotation guidelines are as follows.

1. **Value-judgement** - Does the adjective imply positive value judgment? Most do, including priority claims (e.g., "*first* method to..."). If yes, proceed to steps 2-6. If no, the adjective **not hype**.
2. **Hyperbolic** - Is the adjective hyperbolic or exaggerated? This contains, but is not limited to, a predetermined set of adjectives: *revolutionary*, *unprecedented*, *unparalleled*, *groundbreaking*. If yes, the adjective is **hype**.
3. **Gratuitous** - Does the adjective add little to the propositional content? If removed, and the propositional content and structural integrity of the sentence would remain unchanged (typically when adjective is used in attributive relationship), the adjective is **hype**. If removed, and the propositional content of the sentence would be substantially altered, the adjective is **not hype**.
4. **Amplified** - Is the strength of an adjective amplified through the use of modifiers such as *truly*, *highly*, *completely*? If yes, the adjective is **hype**.
5. **Coordinated** - Is the adjective coordinated with other potentially hype adjectives (e.g., "*innovative* and *creative* researcher")? If yes, then the adjective is **hype**.

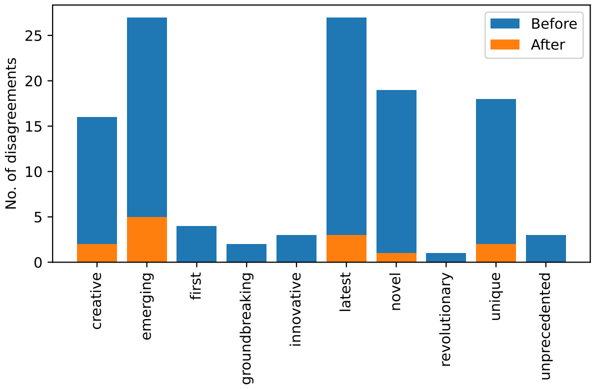


Figure 2: Disagreements between annotators, before and after the discussion. Disagreements were largely resolved, except for *emerging* and *latest*.

6. **Broader context** - When ambiguous, consider whether the sentence contains other instances of potential hype or overt amplification? If yes, the adjective is **hype**.

Guidelines can help a human annotator determine whether an adjective is 'hype', depending on the context. However, in some cases, the guidelines may prove insufficient and require further discussion. Examples of the annotation guidelines are given in Appendix A.

2.2 Annotation process

Authors annotated the dataset separately, without interference. After the initial stage, a discussion session was held to resolve conflicts and reevaluate annotation guidelines.

The initial stage resulted in fair amounts of disagreement, as indicated by the pairwise Cohen's Kappa in Table 1. The disagreement level differs for each adjective; a breakdown is provided in Figure 2. For adjectives corresponding to the **hyperbolic** guideline, the disagreements were minimal; further, *innovative* is rarely disagreed upon, as well as *first*, adjective most commonly used as a numbering device (e.g., *first weeks of therapy*). Adjectives such as *emerging* and *latest* were the most difficult, and the guidelines proved insufficient to fully categorize them. They are often used to refer to *emerging* phenomena when establishing context for the proposed research or the *latest* publications presented at a scientific conference; in these cases, it is required to look at the broader context to determine if their use is highly promotional or not.

Table 1: Pairwise Cohen's Kappa between the annotators A, B and C. Adjusted agreement values after the discussion stage are displayed in brackets.

	A	**B**	**C**
A	–	0.61 (0.94)	0.78 (0.98)
B	0.61 (0.94)	–	0.60 (0.95)
C	0.78 (0.98)	0.60 (0.95)	–

Table 2: Final annotations for each adjective in the dataset.

Adjective	**Hype**	**Not hype**	**Hype %**
creative	33	15	68
emerging	22	23	48
first	17	33	34
groundbreaking	*50*	0	*100*
innovative	41	9	82
latest	28	19	59
novel	18	31	36
revolutionary	*50*	0	*100*
unique	33	15	68
unparalleled	*50*	0	*100*
unprecedented	*50*	0	*100*

After the discussion stage, the initial 119 disagreements were largely resolved. The 13 samples that were not agreed upon were discarded from the dataset, and raised issues regarding the quality of the guidelines, especially for adjectives *emerging*, *latest*, *unique*, and *creative*.

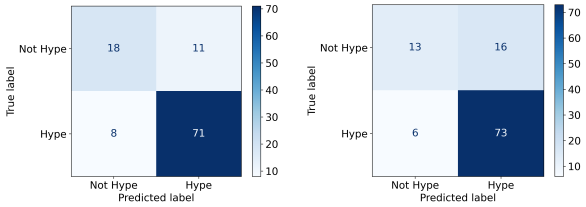
This process yielded a dataset of 537 sentences with potential hype adjectives, 392 deemed as **hype**, and 145 **not hype** by the authors. Some adjectives are more likely to carry promotional intention, while for others, it greatly depends on the context, as seen in Table 2.

3 Preliminary experiments

To understand the difficulty of this task, we conduct preliminary experiments using several traditional text classification methods. Namely, we use Multinomial Naive Bayes (MNB), Multivariate Bernoulli Naive Bayes (MVB), Latent Semantic Analysis (LSA), and Support Vector Machines with a linear kernel (SVC). As features, we consider bag-of-words of unigrams and the averaged word embedding obtained via GloVe[1]. Finally, we obtain a preliminary human baseline.

The dataset was split into a development and hold-out test set in an 8:2 ratio in a stratified manner. A hyper-parameter search was performed on the development set through 10-fold cross-validation, and the performance of

[1] glove.42B.300d



(a) Human (b) GloVe + SVC

Figure 3: Confusion matrices for the human baseline vs. the best performing method. While both had similar Recall on 'hype' sentences, GloVe + SVC struggles to correctly detect 'not hype' sentences.

the best estimator on the hold-out test set is reported. Human baseline was obtained by asking a voluntary researcher to manually go through the hold-out test set; we did not supply them with our annotation guidelines, as we wanted to see what they would deem promotional. Results are in Table 3. Precision, Recall, and F1-score were weighted to account for data imbalance.

All tested methods had similar performances. Interestingly, methods based on BoWwere still capable of performing reasonably well. Furthermore, GloVe + SVC led to the best performance, but it still is behind the human baseline. This result indicates hype detection requires more complex language modeling than BoW features.

A note of caution is due here since our dataset contains about 73% of 'hype' samples. Therefore, these models could just be predicting 'hype' 100% of the time. To better understand this matter, we also analyzed their confusion matrices. For the sake of space, we show only the ones for SVC + GloVe, the best-performing model, and the human baseline in Figure 3. We can see that models struggle to correctly detect 'not hype' sentences, when compared to the human baseline, although GloVe embeddings helped slightly alleviate this issue.

Another point to bear in mind is that given that some adjectives, such as *grundbreaking* and *revolutionary* only appeared in sentences labeled as 'hype', they possibly biased the results towards only classifying sentences as 'hype', as we can observe in Figure 4.

Interestingly, we can see that even though the human baseline did better than the tested methods, its accuracy is not much higher. This result highlights the complexity of this task, showing how hype detection can be tricky even for humans. Therefore, it is likely we need to improve our annotation guidelines for more precise detection.

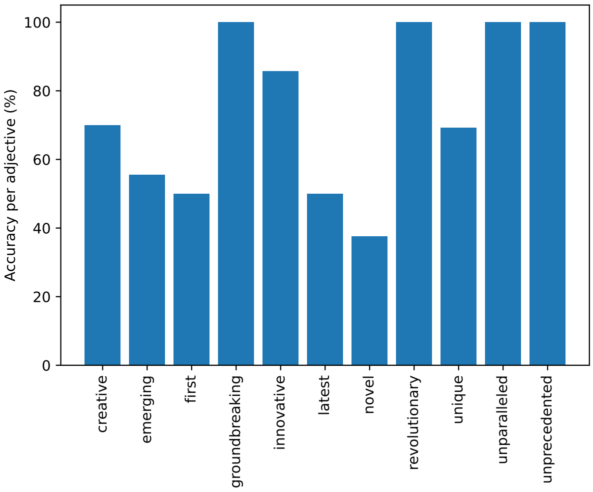


Figure 4: Accuracy per adjective according to MNB + BoW.

Table 3: Performance in terms of Accuracy (Acc.), weighted Precision (Prec.), weighted Recall (Rec.) and weighted F1-score (F1)

Method	Feature	Acc.	Pre.	Rec.	F1
Human	-	0.824	0.819	0.824	0.821
MNB	BoW	0.741	0.713	0.741	0.716
MVB		0.741	0.713	0.741	0.716
LSA		0.685	0.671	0.685	0.677
SVC		0.759	0.736	0.759	0.717
SVC	GloVe	**0.796**	**0.784**	**0.796**	**0.781**

4 Conclusion

To the best of our knowledge, this is the first attempt to reduce subjectivity in identifying promotional language in scientific texts. We developed formal annotation guidelines and applied them to a set of texts from the NIH archive. Multiple machine learning models were used for the defined text classification task, determining whether a sentence containing a potentially promotional adjective is 'hype' or 'not hype'. The results indicate potential but require further feature engineering, a better look at context (via more advanced word embeddings), or the use of contextual models such as LLMs. For future work, we plan on upgrading our guidelines to reflect issues raised in this research, expanding the current dataset with additional annotations before further application and development of classification models.

Acknowledgement

This study was supported by grant No. 21K02919 from the Japan Society for the Promotion of Science. Thanks to Lincon S. Souza for his help in the annotation process.

References

[1] Neil Millar, Francoise Salager-Meyer, and Brian Budgell. “it is important to reinforce the importance of···”: ‘hype’ in reports of randomized controlled trials. **English for Specific Purposes**, Vol. 54, pp. 139–151, 2019.

[2] Neil Millar, Bojan Batalo, and Brian Budgell. Trends in the use of promotional language (hype) in abstracts of successful national institutes of health grant applications, 1985-2020. **JAMA network open**, Vol. 5, No. 8, pp. e2228676–e2228676, 2022.

[3] Christiaan H Vinkers, Joeri K Tijdink, and Willem M Otte. Use of positive and negative words in scientific pubmed abstracts between 1974 and 2014: retrospective analysis. **Bmj**, Vol. 351, , 2015.

[4] Nils B Weidmann, Sabine Otto, and Lukas Kawerau. The use of positive words in political science language. **PS: Political Science & Politics**, Vol. 51, No. 3, pp. 625–628, 2018.

[5] Howard Bauchner. Hype, the responsibility of authors and editors, and the subjective interpretation of evidence. **JAMA Network Open**, Vol. 6, No. 12, pp. e2349125–e2349125, 2023.

[6] Peter Van den Besselaar and Charlie Mom. The effect of writing style on success in grant applications. **Journal of Informetrics**, Vol. 16, No. 1, p. 101257, 2022.

[7] Hao Peng, Huilian Sophie Qiu, Henrik Barslund Fosse, and Brian Uzzi. Promotional language and the adoption of innovative ideas in science. **Proceedings of the National Academy of Sciences**, Vol. 121, No. 25, p. e2320066121, 2024.

[8] Kristen Intemann. Understanding the problem of “hype”: Exaggeration, values, and trust in science. **Canadian Journal of Philosophy**, Vol. 52, No. 3, pp. 279–294, 2022.

[9] Neil Millar, Bojan Batalo, and Brian Budgell. Promotional language (hype) in abstracts of publications of national institutes of health–funded research, 1985-2020. **JAMA Network Open**, Vol. 6, No. 12, pp. e2348706–e2348706, 2023.

[10] Neil Millar, Bojan Batalo, and Brian Budgell. Trends in the use of promotional language (hype) in national institutes of health funding opportunity announcements, 1992-2020. **JAMA Network Open**, Vol. 5, No. 11, pp. e2243221–e2243221, 2022.

[11] Andrew Hardie. Cqpweb—combining power, flexibility and usability in a corpus analysis tool. **International journal of corpus linguistics**, Vol. 17, No. 3, pp. 380–409, 2012.

A Some examples in the annotation guidelines

Guideline 1: Value-judgement. Does the adjective imply positive value judgment?

- **YES** - Most adjectives will imply a value judgement. This includes priority claims:
  - *Our study will be the* ***first*** *to ...*
- **NO** - Typically, acronyms, technical/domain-specific meaning, and literal meaning:
  - *To aid these efforts,* ***Creative*** *Scientist, Inc. (CSI)...*
  - *Our curriculum emphasizes the development of critical and* ***creative*** *independent thinking...*
  - *In the* ***first*** *aim we test the hypothesis...*

Guideline 2: Hyperbolic. Is the adjective hyperbolic or exaggerated?

- **YES** - A relatively unambiguous class that can (likely) be pre-determined:
  - *revolutionary; unprecedented; unparalleled; groundbreaking*

Guideline 3: Gratuitous. Is the adjective gratuitous, adding little to the propositional content?

- **YES (1)** - If removed, the propositional content and structural integrity of the sentence would remain basically unchanged (typically when adjective used in attributive relationship).
  - *To address this, we developed 2* ***innovative*** *technologies.*
  - *Delivering SGR interventions via text messaging is an* ***innovative*** *way to increase the reach of this cessation intervention...*
- **YES (2)** - Represents a tautology or is redundant?
  - *discovered a* ***novel*** *gene...*
- **NO (1)** - If removed the propositional content of the sentence would be substantially altered.
  - *This is a high risk and high impact project that uses a* ***novel*** *approach to aggressively treat local - regional disease.*
- **NO (2)** - The sentence gives justification for the claim (typically when adjective used in predicative relationship).
  - *The proposed study is* ***innovative*** *because no previous research has identified how MBC...*

Guideline 4: Amplified. Is the strength of the adjective amplified?

- **YES** - The strength of the adjective made stronger through the use of modifiers:
  - *truly novel; highly innovative; completely unique; etc.*

Guideline 5: Coordinated. Is the adjective COORDINATED with other hype candidates?

- **YES** - Adjective is co-ordinated with one or more hype candidates (adjective stacking):
  - *...****innovative*** *and* ***creative*** *leader...*
  - *...****creative****,* ***collaborative****, and culturally* ***diverse*** *translational scientists...*

Guideline 6: Broader context. When ambiguous, consider whether the sentence contains other instances of potential hype or overt amplification.

- *This* ***transformative*** *work will be the* ***first*** *study to achieve this level of*
- *The faculty has an* ***outstanding*** *track record of* ***creative*** *and high - profile research ,* ***superb*** *mentoring , and* ***robust*** *research funding , and thus attracts* ***outstanding*** *trainees*