A probabilistic model of 'Hype' in scientific abstracts.

Apratim Mishra[1], Jana Diesner[2] and Vetle I. Torvik[3]

[1]*apratim3@illinois.edu*
School of Information Sciences, University of Illinois Urbana-Champaign, 501 E. Daniel Street, Champaign, IL (United States of America)

[2]*jdiesner@illinois.edu*
School of Information Sciences, University of Illinois Urbana-Champaign, 501 E. Daniel Street, Champaign, IL (United States of America)

[3]*vtorvik@illinois.edu*
School of Information Sciences, University of Illinois Urbana-Champaign, 501 E. Daniel Street, Champaign, IL (United States of America)

Abstract
Hype in scientific writing is purportedly on the rise. Prior studies have used the presence of emotive words such as '*novel'*, '*outstanding'*, and '*unique'*, as binary indicators of hyperbolic or promotional language (i.e., hype). In this study, we propose a probabilistic model of 'hype' based on the context of the candidate emotive word, and a measure of its overall propensity towards non-neutrality. Analyzing 44 previously proposed 'hype' words across 18.8 million PubMed abstracts, we find that the majority of instances appear neutral, for example, when they are part of a phrase pointing to a technical concept (e.g., *major* histocompatibility complex). Certain words such as '*promising'* and '*encouraging'* have a higher propensity for hype, whereas others have low, such as '*major'* and '*novel'*. The model provides a more precise (probabilistic) labeling of scientific abstracts which should enable further study of 'hype' and its role in scientific communication.

Keywords: Text mining, Sentiment analysis, PubMed, Hype, Probabilistic Modelling

Introduction
In recent years, there have been several reports stating that authors of scientific papers deliberately use subjective language to overstate the importance of their results or embellish methods or results to appeal to readers (Martin, 2009; Vinkers, 2015). An editorial in Nature Medicine (2006) specifically pointed out that numbers should speak for themselves, and the integrity of numbers or data has become a thorny issue. This word usage is criticized as dramatic language by editors and deemed unethical to utilize for the sole purpose of encouraging positive appraisal (Jones, 2017; Wheatley, 2014). Several studies in diverse fields such as physics (Najjar, 1987), biology (Huckin, 1995), and computer science (Perez, 2014) have indicated that these disciplines have used persuasive promotional rhetoric.

In a corpus study of 400 'hype' items over 50 years in four disciplines, Hyland (2021) traces a shift from a neutral stance in research papers, with the most increase in the hard sciences, especially biology. In a recent study of Covid -19 research, Hyland, and Jiang (2021) also showed a significant increase in hype to stress on paper novelty, potential, and overall contribution. While it has been recognized that authors do work to promote their research (Najjar, 1987), studies have faulted researchers for making their findings appealing and more groundbreaking than they actually are and overstating or exaggerating potential implications (Ecklund, 2015). Hyperbolic words in academics have been used to glamorize results, secure grants and get published; they have received several names such as 'marketization' (Fairclough, 1993), 'elements of selling' (Bhatia, 1993), 'boosterism' (Swales, 2004), 'boosting' (Hyland, 2012), 'quasi advertising discourse' (Lindeberg, 2004), 'linguistic spin' (Lazarus, 2015), 'hype and 'value-laden vocabulary'(Martin, 2009; Miller, 2019). Promotional language has also been assessed within unintentional or intentional '*spin'*, where authors have identified strategies to

mislead reporting, provided inadequate interpretation, or extrapolate results to modify reader interpretation (Lazarus, 2015). In biomedical data, the absolute frequency of positive words, such as 'novel', 'robust', and 'innovative' has increased from 2% to 17.5% from 1974 – 2014, which is a relative increase of 880% over four decades (Vinkers, 2015). As academia has become more competitive over time, publishing has become a method of monetization and has increased its importance in-for profit societies (Johnson, 2018). Therefore, there is a need to provide a more nuanced view of scientific 'hype'.

Specifically, we present our work covering the following issues:

- While studies conducted point out the significant increase in the use cases of 'hype' words, previous work has provided limited context for their relative use
- We control for word phrases such as '*novel_mutation'*, '*central_portion*', and '*major_histocombatibility',* which either present a methodological viewpoint or are common due to word pair frequencies. We also control for negations and limit the scope of hype for positive contexts.
- 'Hype' is intentional promotion; therefore, we separate the intended use of 'hype' into 'present hype', which should include language promoting the papers' own results and 'past hype', that is language referencing background data.
- Finally, we present a mixture model covering word usage based on abstract percentile position which approximates 'past hype', background noise, and 'present hype' for a paper.

Dataset

To identify potential cases of 'hype', we consider 44 unique words, collected from several prior studies (Jones, 2017; Wheatley, 2014; Vinkers, 2015), and locate all of their instances across 18.8 million PubMed abstracts after lowercasing and lemmatizing words. The initial count for the ten most common 'hype' words is: '*major'* (435076), '*novel'* (273000), '*central* (212101), '*strongly* (180674), '*critical* (162638), '*markedly* (129,551), '*unique* (119199), *excellent* (77607), '*crucial* (63676), '*promising'* (58055).

Data and Code presented: https://github.com/apratim-mishra/ISSI2023/.

Method

We interpret the position of the word in the abstract based on the conventional IMRaD (Nair, 2014) structure of scientific writing.

- Introduction: This section contains background information and motivates the problem addressed. This section captures 'past hype' which comprises hyperbolic language about prior work or the importance of the problem.
- Methods: This section describes the research design. 'Hype' words occurring here are often part of technical phrases that are not hyperbolic, such as '*novel mutation'*, '*vital status,* and '*supportive care'*.
- Results: This section presents findings and outcomes, and hyperbolic language here represent 'present hype'.
- Discussion: This section discusses the results and their potential implications, and hyperbolic language here represents 'present hype'. It should have the greatest propensity of 'hype'.

Contextualizing hype

While researchers do not have an exhaustive list of possible 'hype' words as well as the relative extent of 'hype' with which each of the words dramatizes or promotes a paper's content, there has also been limited research into considering the subject towards which the hype is designated. Therefore, to improve the contextualization of 'hype', we theorize that the abstract position of the 'hype' word usage is necessary for better comprehension.

A word used in the earlier part of the abstract would be used to describe the background works that present the readers with a framework for the contents of the paper. While the percentile cutoff for the introductory part of the abstract is hard to designate, researchers using 'hype' words in the early parts of the abstracts are referring to earlier works and using language and linguistic principles that related papers have used.

E.g., *"Prevention of variceal bleeding, a major cause of morbidity and mortality, is an important goal in the management of patients with portal hypertension (PHT)." (PMID: 10197489)*

This instance is the very first sentence of the abstract of the biomedical literature, which describes describing the backdrop setting, which is not the contribution of the paper. In our scenario, these would not be expected instances of 'hype'. They are laying the foundation of their subject but are referring to another research.

Additionally, if there are cases where the specific word fails to modify the knowledge claims posited by the paper, they should be exempted from the tabulation of 'hype'. These should typically be among the common instances of the word n-grams that are uniformly distributed across the IMRaD abstract sections. For e.g.,

"treating peripheral blood mononuclear cells with agents inhibiting non-major histocompatibility complex-restricted cytotoxic activity" (PMID: 1909874)

The statement is a case of a methodological concept being defined; '*major histocompatibility'* is not used as an adjective for promotion but as part of a common use bigram. The next example is a case of negation which is another type of phrase filtering.

"The morphological changes were not remarkable in the liver cells throughout the study" (PMID: 1875893)

However, phrases describing results are examples of 'hype' where the content of the paper in question is being hyped. For e.g.,

"In Swiss mice, animals with high natural resistance to hepatitis virus, the pathogenicity of this agent was markedly enhanced by combined infection with eperythrozoa." (PMID: 13109101)
Similarly, discussing of possible implications of the research would be an example of 'hype' in the discussion the paper results. For e.g.,

"However, the ability of stress protein induction to protect against injury from glutamate may offer a novel approach toward ameliorating damage from excitotoxins". (PMID: 1764242)

This example of the word '*novel'* points out possible implications that the paper offers in the last statement of the abstract, providing a concluding remark of the paper's implications.

Modelling word distribution

Almost all 'hype' words exhibit a bimodal distribution across abstracts with the two modes corresponding to the IMRaD sections, Introduction and Results/Discussions, respectively. Therefore, our model contains 3 components: one for each of the two modes which represent 'hype' (i.e., signal), and a uniform baseline that represents neutral instances (i.e., noise). This model permits the computation of a probability of 'hype' by the signal proportion at any given position of the 'hype' word in the abstract.

Each of the two modes is captured with a logistic or an exponential component and the baseline is captured by a uniform distribution as follows:

Equation 1: Model Hype Word Distribution

$$Pr\{position = x\} = e + q\frac{1}{1+e^{-ax-b}} + w\frac{1}{1+e^{-c(x-1)-d}}$$

Where, 0<x <i for the first sigmoid function, j<x<1 for the second sigmoid function, and i<j
And q, w, and e >0, and q + w + e = 1

The model includes 8 free parameters: a, b, q, and i characterize 'past hype', c, d, w, and j characterize 'present hype', and e the 'noise'. The best fitting parameters are obtained using a grid search minimizing the chi-squared statistic comparing the observed versus the expected counts across 30 equally spaced bins, excluding the first and last bins. The overall probability of past and present 'hype' would be q and w respectively. We aim to only model research papers that follow the IMRaD structure; we filter out PubMed article types such as case reports, editorials, letters, biographical, etc.

Figure 1 illustrates the model fit for 2 selected words for the '*Before'* model. The two distributions follow the same shape as specified in the model, however, '*promising'* has a much bigger signal compared to '*novel'*. We also note the edge effects (first and last bin deviate from the model) that are likely due to the unequal position in any particular sentence.

While modelling, we observe that for a minority of words, this bimodal structure is not followed, such as, for words like '*markedly'*, and '*strongly'*. These words are currently excluded from our analysis, as such words with strong peaks in the 'Results' section do not fit our model.

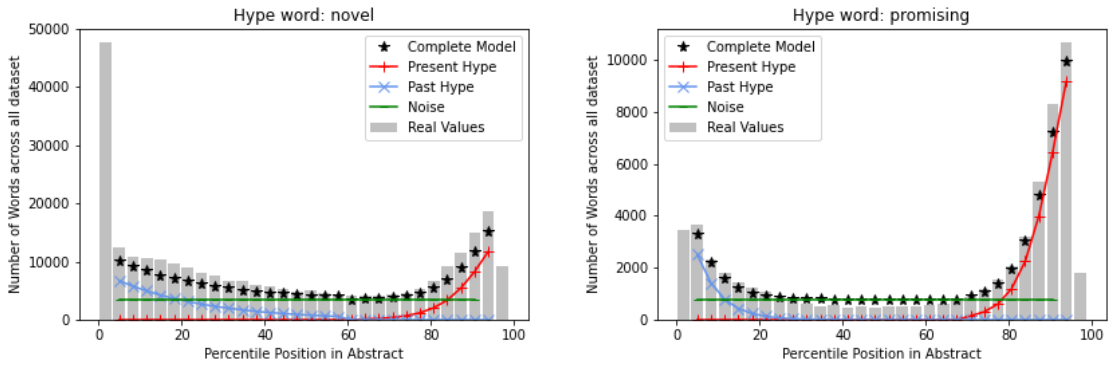

Figure 1: Positional distribution of 2 selected 'hype' words. Grey represents observed counts, black represents the model fit.

Next, we also present an '*After'* version of the fitted model to improve the signal for hype propensity. Firstly, we control for phrases including negations, for e.g., that start with no, nor, not, non, and phrases that significantly reduced the noise component, 'e'. We have also removed technical phrases from the UMLS that bring biomedical terminologies or phrases that are very cohesive (Torvik, 2007). A high cohesive score indicates a greater tendency to be related to a knowledge concept, referred to as 'MeSH' terms, for e.g., '*essential amino'* is a technical term signifying an important concept rather than hyping a narrative.

Table 1 shows the before and after 'noise' proportion of a set of words. It shows that not all words have an equal amount of signal, and that 'noise' is significant for some of them. Table 2 shows the estimated probability of 'present hype' for the 90th percentile position which is expected to be a high value.

Table 1: Overall Noise factor for some words

	Major	Novel	Central	Critical	Unique	Excellent	Crucial	Promising
Before	0.91	0.556	0.511	0.515	0.9	0.47	0.50	0.427
After	0.913	0.557	0.513	0.517	0.89	0.47	0.49	0.426

Table 2: Present Hype Probability at the 90th percentile abstract position

	Major	Novel	Central	Critical	Unique	Excellent	Crucial	Promising
Before	0.077	0.77	0.512	0.76	0.63	0.761	0.41	0.920
After	0.098	0.78	0.53	0.77	0.639	0.762	0.425	0.923

While some words such as '*major'* and '*novel'* have a far greater count in abstracts, the extent to which they 'hype' the paper content is lower compared to others and is contingent on the word location. A larger total noise factor indicates a flatter representation of the 'hype' term indicating the word usage in wide varying contexts. Whereas a word like 'promising' is used to large margin in the discussion section, hyping the paper implications.

Discussion

We notice that the distributions for each of the 'hype' words vary a lot. With the fitted model, we can find out the hype probability for each abstract percentile. This can be the basis for the relative nature of hyperbole among the different candidate 'hype' words. A limitation of the model is the IMRaD assumption; the effect on the model is that the noise component might be higher. Ideally, if the IMRaD assumption is violated, for e.g., if the sections have a different order or if the abstract is unstructured, the estimated probabilities based on position would be inaccurate. Ideally, one would label a sentence in the IMRaD sections without consideration of the position.

This method can help in modeling the 'hype' for a new paper not present in the dataset, based on the presence of one or more 'hype' words and their abstract position. The mixture model provides a good fit for the dataset and presenting 'hype' as a probability offers contextualization to the embellishment in scientific literature. The usefulness of the model is based on its good fit and its simplicity, as compared to NLP methods of querying relations in a biomedical domain. This can be correlated to downstream tasks such as citation impact as well as variables associated with author collaboration.

An important issue is the problem of 'word sense disambiguation', while all studies reflect the word '*outstanding'* as hype, it is rarely used as one; it is commonly used to dentote 'unpaid' or 'owed' work, rather than as a 'distinguished' achievement. Finally, most importantly, our model is based on the IMRaD assumption, which may not hold true for all scientific abstracts.

References

Bhatia, V. (1993). Analyzing genre: Language use in professional settings. Longman.

Lazarus, C., Haneef, R., Ravaud, P., & Boutron, I. (2015). Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. (pp. 1-8). BMC Medical Research Methodology, 15 (1).

Vinkers, C. H., Tijdink, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: Retrospective analysis. (pp. 1-6). BMJ, 351.

Elaine, H.E., Johnson, D.R., & Matthews, K.R.W (2015). Commentary: Study highlights ethical ambiguity in physics. (pp. 8-10). Physics Today, 68 (6).

Fairclough, N. (1993). Critical discourse analysis and the marketisation of public discourse: The universities. (pp. 133-168). Discourse & Society, 4 (2).

Fyfe, A., Coate, K., Curry, S., Lawson, S., Moxham, N., Rostvik, C.M. (2017). *Untangling Academic Publishing: a History of the Relationship between Commercial Interests, Academic Prestige and the Circulation of Research.* Retrieved from https://doi.org/10.5281/zenodo.546100

Huckin, C. B. (1995). Genre Knowledge in Disciplinary Communication. In *Lawrence Erlbaum.* , Hillsdale, New Jersey.

Hyland, K. (2000). Disciplinary discourses: Social interactions in academic writing.

Hyland, K. (2012). "The past is the future with the lights on": Reflections on AELFE's 20th birthday. (pp. 29-43). Iberica, 24.

Jiang, K. H. (2021). The Covid infodemic: competition and the hyping of virus research. Int. J. Corpus Linguist. (2021).

Jones, S. S. (2017). *Superlative scientific writing.* ACS Catalysis.

Hyland, K., & Jiang, F.K. ((2021). 'Our striking results demonstrate …': Persuasion and the growth of academic hype. *Journal of Pragmatics, Volume 182,*, pp. 189-202.

Lachowicz, D. (1981). On the use of the passive voice for objectivity, author responsibility and hedging in EST. *Science of Science, 2(6),* 105-115.

Lerchenmueller , M.J., Sorenson, O., & Jena, A.B. (2019). Gender differences in how scientists present the importance of their research: observational study. BMJ.

Lindeberg, A. (2004). Promotion and politeness: Conflicting scholarly rhetoric in three disciplines. Abo Akademi University Press .

Martin, V. F. (2009). Marketing data: Has the rise of impact factor led to the fall of objective language in the scientific article? *Respiratory Research*, 1-5.

Millar, N., Salager-Meyer, F., & Budgell, B.. (2019). "It is important to reinforce the importance of…": 'Hype' in reports of randomized controlled trials. (pp. 139-151). English for Specific Purposes, 51 .

Nair, P.R., & Nair, V. (2014). Scientific Writing and Communication in Agriculture and Natural Resources. . Cambridge International Law Journal.

Najjar, J. S. (1987). The writing of research article introductions. Writ. Commun., 2 (4) .

Perez, M.A. (2014). Convincing peers of the value of one's research: a genre analysis of rhetorical promotion in academic texts. In *Engl. Specif. Purp., 34 (2014)* (pp. 1-13).

Johnson, R., Watkinson, A., & Mabe, M. (2018). The STM Report. Holland: International Association of Scientific, Technical and Medical Publishers,.

Swales, J. (2004). Research genres: Explorations and applications. Cambridge University Press.

(2006). *Truth in Numbers.* Nature Medicine.

Torvik, V.I., & Smalheiser, N. R. (2007). A quantitative model for linking two disparate sets of articles in MEDLINE. *Bioinformatics, Volume 23, Issue 13*, 1658-1665.

Wheatley, D. (2014, February). Drama in research papers. European Science Editing.