arXiv vs. snarXiv

Tyler Blanton and Sam Kowash
University of Washington, Department of Physics

Background

The arXiv is a popular e-print repository for publications in physics, astronomy, and other quantitative sciences. It hosts nearly 1.5 million papers, of which ~120,000 are in theoretical high-energy physics (`hep-th`). Physicist David Simmons-Duffin developed a program known as the snarXiv, which procedurally generates abstracts in the style of `hep-th` from a context-free grammar.

Humans (even physicists) have surprising difficulty determining whether a given abstract is from the arXiv or the snarXiv. Across 750,000 guesses collected through an online sorting game, players succeeded at picking the genuine paper from a pair only 59% of the time. Try it for yourself below!

Abstract A

In the 20th century, a fair amount of work was done demystifying QED in the presence of a stack of canonical co-isotropic branes. In this paper, we make contact with analyzing heterotic strings, consequently reconstructing perturbation theory on $\mathbb{C}^n$, and classify anomalous dimensions in loop models with sleptons. Our computation of the solution of magnetic dualities in models of hadrons provides a certain notion of perturbation theory (taking into account cosmic rays at Λ_{QCD}). Our results prove that decay constants turn out to be equivalent to an instanton at the Planck scale. Finally, we establish that sleptons can be brought to bear in reformulating heavy ions.

Abstract B

We study the effective action of the heterotic string compactified on particular half-flat manifolds which arise in the context of mirror symmetry with NS-NS flux. We explicitly derive the superpotential and Kähler potential at lowest order in α' by a reduction of the bosonic action. The superpotential contains new terms depending on the Kähler moduli which originate from the intrinsic geometrical flux of the half-flat manifolds. A generalized Gukov formula, valid for all manifolds with SU(3) structure, is derived from the gravitino mass term. For the half-flat manifolds it leads to a superpotential in agreement with our explicit bosonic calculation. We also discuss the inclusion of gauge fields.

We explored computational approaches to this classification problem to investigate whether its difficulty stems from snarXiv's genius, or humans' bewilderment in the face of unfamiliar jargon. (Explore snarXiv at: `https://github.com/davidsd/snarxiv`.)

n-gram model

One approach to text classification is to develop a *language model.* Given a class $Y \in \{-1, 1\}$ (with the negative sign referring to arXiv and the positive to snarXiv) and a document X consisting of words $\{w_i\}_{i=1}^N$, we want to characterize the probability

$$\mathbb{P}(X|Y) = \prod_{i=1}^{N} \mathbb{P}(w_i | w_1^{i-1}, Y),$$

where w_1^{i-1} is the sequence of the first $i-1$ words. The sample space dwarfs the available data and this can never be satisfactorily trained. Instead we assume that a word's probability depends only on the preceding $n-1$ words, reducing the training task to estimating the probabilities of so-called n-grams from their frequencies in the training corpora.

$$\hat{\mathbb{P}}(X|Y) \equiv \prod_{i=1}^{N} \hat{\mathbb{P}}(w_i | w_{i-n+1}^{i-1}, Y) \equiv \frac{C_Y(w_{i-n+1}^{i}) + 1}{C_Y(w_{i-n+1}^{i-1}) + V},$$

where $C_Y(w_{i-n+1}^{i})$ is the number of times that n-gram occurs in the corpus for class Y and V is the size of our vocabulary. In our experiments we studied 1-grams (known as the bag-of-words model) and 2-grams.

Naive Bayes

The simplest classifier based on n-gram frequencies sorts a document into whichever class has a higher posterior probability based on our estimate:

$$\hat{Y} = \underset{Y \in \{-1,1\}}{\arg\max}\, \mathbb{P}(Y) \hat{\mathbb{P}}(X|Y),$$

where $\mathbb{P}(Y)$ is the fraction of training abstracts in class Y.

Likelihood-ratio test

The LR test improves upon naive Bayes by giving us a hyperparameter η to adjust the sensitivity of the test. Given the ratio $\Lambda(X) \equiv \hat{P}(X|Y=1)/\hat{P}(X|Y=-1)$, the classifier δ_{LR}^* chooses class 1 (snarXiv) with probabilities

$$\mathbb{P}(\delta_{LR}^*(X) = 1) = \begin{cases} 1, & \Lambda(X) > \eta \\ 1/2, & \Lambda(X) = \eta \\ 0, & \Lambda(X) < \eta \end{cases}.$$

This reduces to the previous case when η is the ratio of arXiv abstracts to snarXiv in the training set.

tf–idf

Another approach is to map documents into a continuous vector space and apply typical statistical learning methods. We used a tf–idf method, where each document gets a vector d with an entry d_i for each word in the vocabulary, weighted by

$$d_i = \mathrm{tf}_{i,d} \cdot \mathrm{idf}_i,$$

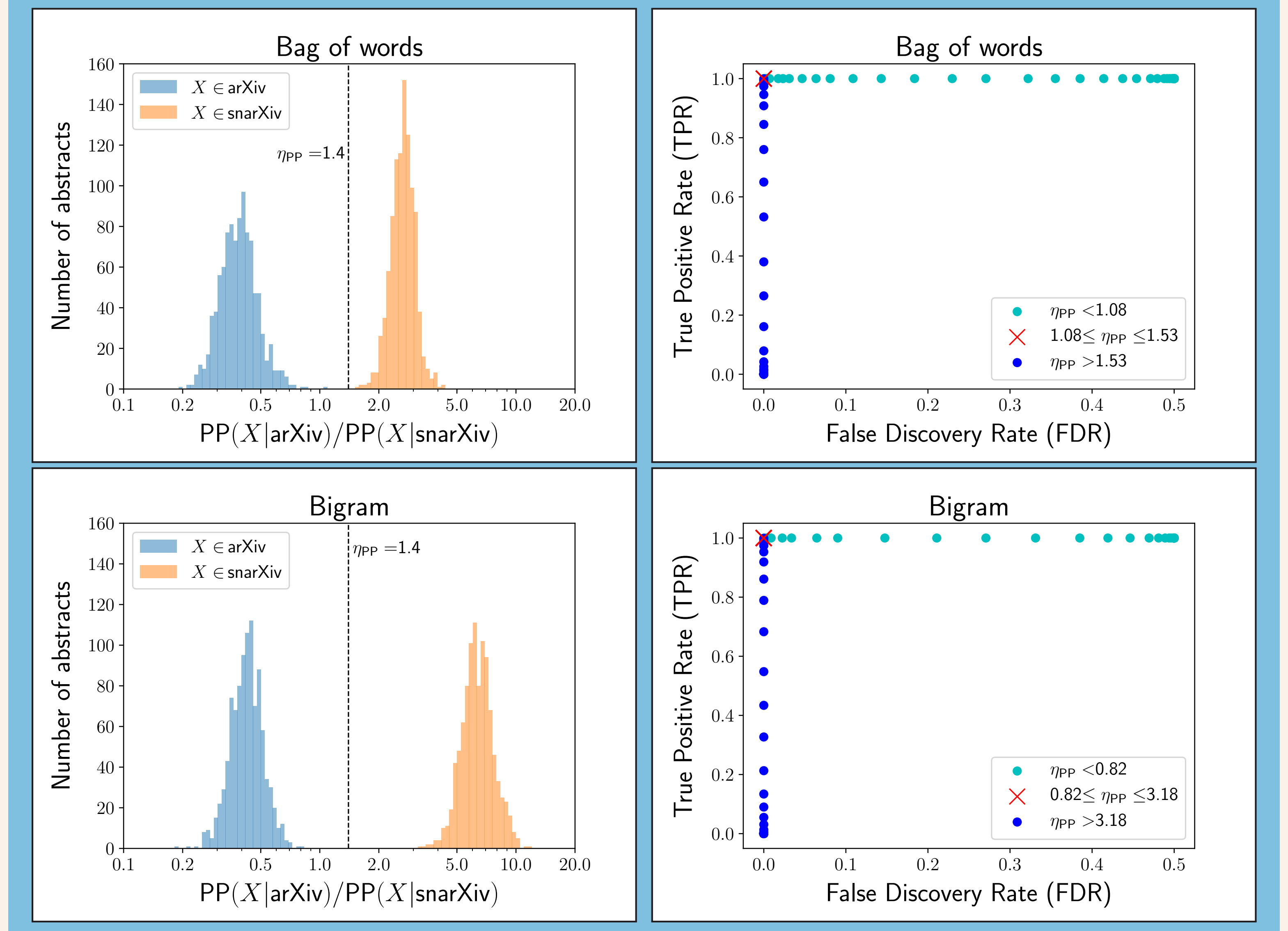
where we define

$$\mathrm{idf}_i \equiv \log_{10} \frac{N+1}{\mathrm{df}_i + 1}, \quad \mathrm{tf}_{i,d} \equiv \begin{cases} 1 + \log_{10} C_{i,d} & : \quad C_{i,d} > 0 \\ 0 & : \quad \text{else} \end{cases},$$

in which $C_{i,d}$ is the number of times w_i appears in the document and df_i is the number of documents containing w_i. This weighting scheme is designed to focus on words which are less common in the training set, and ignore words like "the" and "of" which lack discriminating power. We used these vectors as features in an L^2-regularized logistic regression to classify abstracts.

Results

Naive Bayes classifier: TCA $\approx$ 99.94% (TPR = 100%, FDR $\approx$ 32%)
tf–idf + logistic regression: TCA $\approx$ 99.96% for $\lambda = 2.5 \times 10^{-8}$
Likelihood-ratio test: TCA $\approx$ 99.98% for $\eta_{\mathrm{PP}} = 1.4$



For each model, we trained on 1,000 abstracts of each set and tested on 4,000 of each, using a vocabulary of 15,315 unique words that appeared at least twice in a separate training corpus of 12,000 abstracts from each set. The naive Bayes classifier (using bag-of-words) is the worst performer, but still boasts a total classification accuracy (TCA) of 99.4%, with the only misclassifications being false positives (arXiv mistaken for snarXiv).

The likelihood-ratio test (for both bag-of-words and bigrams) and logistic regression (on tf–idf) perform even better, each sporting a TCA greater than 99.95%. Comparing the bag-of-words and bigram models in the likelihood-ratio test, we find that while both of the resulting classifiers are able to achieve a TCA of at least 99.98% for appropriate choices of η_{PP}, the range of η_{PP} which achieve these accuracies is significantly wider in the bigram model; in other words, adding context to words makes the contrast between arXiv and snarXiv abstracts much more pronounced.

Conclusions

Ultimately, we found that distinguishing arXiv from snarXiv is not just easier for computers, but *trivial.* As we see in the histograms, the populations are statistically well-separated even in BoW analysis, but strikingly so when 2-grams introduce context. We expect this to sharpen with higher context, as the snarXiv produces long-range coherence only by random chance, and its juxtapositions of unrelated concepts will become more evident.

References

Dan Jurafsky and James H Martin. *Speech and language processing*, volume 3. Pearson London, 2014