Modeling "Newsworthiness" for Lead-Generation Across Corpora

ACM Reference Format:
. 2019. Modeling "Newsworthiness" for Lead-Generation Across Corpora. 1, 1 (December 2019), 5 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Journalism is the identification and publishing of interesting pieces of information, i.e., information that is *newsworthy*. "Newsworthiness", classically, refers to information that informs voters in a democracy. Formalizing a ranking of content based on newsworthiness is essential to helping us identify democratically relevant information.

However, newsworthiness is ill-defined. A piece of information might be newsworthy at one time, but less at others, or may become newsworthy depending on subsequent events. Judging the newsworthiness of a piece of information requires intensive human efforts, based on intuition about what kind of information is important for voters.

To this end, we offer a narrower, operational definition of "newsworthiness" that seeks to interpret and apply historical expert judgements. Our definition is: *how likely is this piece of information to appear on the front page of a major newspaper?* With this definition, we propose a simple classification task: is this record similar to news articles that have appeared on the front page?

In this work, we train models to learn "newsworthiness" by classifying the page that newspaper articles are published on. We use these models to sort documents in other corpora used by journalists. Our core contributions are: (1) a re-formulation of the "newsworthiness" definition and its problem setup, (2) a demonstration it can be accurately modeled with modern language models, and (3) an analysis of the decision-making processes in the models we use.

We see potential applications for such predictions to be incorporated in algorithmic ranking systems throughout the web. Additionally, we see such an approach helping journalists filter information. Journalists use various corpora – e.g., court cases, city council minutes – in their day-to-day work to keep abreast of the workings of government, and such models, we show, can surface relevant documents for journalists.

2 RELATED WORK

One challenge in computational journalism is *lead generation*, or identifying and surfacing interesting information that could spark story ideas, or leads [2].[1]

[1] *Computational journalism* is an emerging field aimed at identifying applications of statistical and computational approaches to impact the traditional journalistic practice.

Author's address:

XXXX-XXXX/2019/12-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Corp.	Top predictions from LR
Meeting Min.	*Rules* which prohibit use of funds for 2026 World Cup unless U.S. Soccer Fed. provides equitable pay to U.S. Women's and U.S. Men's Team
State Bills	*Bill* requiring school districts to participate in Medicaid for health and social services.
Appeals Courts	*Indictment* returned against Governor Rick Perry for ... exercising authority to veto appropriations vested in the Governor by Texas Constitution.

Table 1. Sample of top "newsworthy" records from various corpora produced by our models.

One approach to lead-generation is to quantify contents' *relevance* to a given topic, its *uniqueness*, and its *sentiment*. [4] present metrics for identifying such content and applies it to surfacing tweets. This approach is useful for generating leads for events journalists know to search for, like political speeches, but are limited in identifying new topics of coverage. Our approach does not place such a constraint on journalists, and can surface novel content independent of preconceived topic. A second approach is *anomaly detection*. Systems like Newsworthy analyze open-source, numerical datasets, like polling data and housing market data, to discover outliers [5], which are surfaced to journalists to investigate. This approach might be relevant for data-driven stories, but would not capture many event-driven stories. Our approach surfaces textual data, thus operates in a different domain. A third approach involves fact-checking: systems like ClaimBuster scan news, speeches and social media for claims being made by politicians. Once claims are identified, they are forwarded to journalists to check [1].All three of these approaches utilize specific, expert-designed metrics for newsworthy content. Our approach sidesteps these systems and directly models historical newsworthiness.

3 PROBLEM DESCRIPTION

Our goal is to model p(newsworthy|text) for any input text. So, we consider two sets of corpora: a set of labeled corpora $C^{'} = \{c_i^{'}\}$ and a set of unlabeled $C^{''} = \{c_k^{''}\}$. For each article $a \in C^{'}$, we have labels for newsworthiness (whether the article was published on the front page or not), for document $d \in C^{''}$, we do not.

Our predictive tasks are as follows: we seek to build models that (1) accurately classify labels on $C_{.}^{'}$, which we evaluate on a held-out set. And (2) generalizes to $C_{.}^{''}$, which we will evaluate using expert annotation.

4 DATA

In order to set up the problem described in Sec. 3 section, we collect four corpora: one labeled and three unlabeled. We are careful to avoid the following scenario: a document in $C^{''}$ is directly used to report for an article published in $C^{'}$, leading our model to overfit specific topics or people. So, we choose date-ranges for $C^{'}$ and $C^{''}$ such that no $a \in C'$ could reference $d \in C''$.

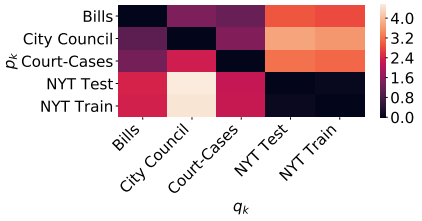

Fig. 1. Variation between the corpora, shown with KL-Divergence over unigram counts for each corpus. The Y-axis shows the base distribution, p_k and the X-axis shows the comparison distribution, q_k.

$C_1^{'}$**: New York Times Annotated Corpus (1987-2007)**[2] We use the *New York Times* Annotated Corpus as training and evaluation, which contains 1.8 million articles published from 1980–2007, (40, 000 front-page). Each article has a number of attributes, including those relevant for our purposes: headline, full-text and page-number.

$C_2^{''}$**: Los Angeles City Council Meeting Minutes (2015-2019)**[3] The Los Angeles City Council publishes summaries of topics discussed in council meetings online. We scrape all city council meetings occurring between 2015-2019. For each meeting, we parse the separate agenda items, which consist of a title, a case-number, and a brief description of the item. We collect 113, 000 documents.

$C_3^{''}$**: State-level Bills (2010-2018)**[4] We collect all state-level bills passed between 2010-2018, as recorded by Open Secrets. In total, we collect 1.04 million documents. The information provided includes title, a brief description, and subject-tags.

$C_4^{''}$**: Opinions from Appeals Court Cases. (2008-2018)**[5] We use CourtListener, an open-source scraper that compiles court-dockets from different state and federal court houses, to collect all documents produced by appeals cases, totaling over 110, 000 documents. For each document, we have the full-text of the opinions offered, including defense and judgement.

Our corpora exhibit a range of variance, as shown in Figure 1. The train/test splits, discussed in the next section, show a low KL-divergence of .1 in both directions. The highest divergence is observed between D_{kl}(“NYT Test”|| “City Council”) = 4.5. (Our $|Vocab| = 21,000$)) – City Council is the most local of our corpora.

5 EXPERIMENTAL RESULTS

We preprocess all of our corpora to eliminate a list of stopwords specific to newspaper publishing (ex: “op-ed”, “sportsmonday”, “business review”).[6] We train each of our models listed in Table 2 on an artificially balanced training set from $C_1^{'}$ from the years 1987-2001 ($y_1^{train} = y_0^{train} \approx 45,000$ articles).

Goal 1: Performance on $C^{'}$ **heldout** To test if our models accurately predict newsworthiness, we evaluate on an unbalanced training set from $C^{'}$ from the years 2001-2007 ($y_1^{test} \approx 11,000$ articles, $y_0^{test} \approx 210,000$ articles.) As shown in Table 3, the top-scoring model is RT (.93 AUC).

[2]https://catalog.ldc.upenn.edu/LDC2008T19
[3]https://cityclerk.lacity.org/lacityclerkconnect/
[4]https://openstates.org/
[5]https://www.courtlistener.com/
[6]We identify this list through iteratively training LR and examining top coefficients.

1. LogReg. [8]:	**LR**
2. FastText v1.0 [6]:	**FT**
3. BERT-Base [3]:	**BT**
4. RoBERTa-Base [7]:	**RB**

Table 2. Classification approaches tested on $C^{'}$.

Method	**LR**	**FT**	**BT**	**RT**
AUC	.85	.88	.91	.93

Table 3. Eval. 1: AUC on $C^{'}$, all models.

Table 3: Words receiving top attention weights to the *[CLS]* token in the final layer BT.

Front Page		**Not Front Page**	
Word	Atten.	Word	Atten.
threatens	.059	suffolk	.078
startling	.053	diary	.060
follows	.053	connecticut	.053
stunned	.052	knicks	.049

Table 4: Top attention weights to the *[CLS]* token in the final layer RT.

Front Page		**Not Front Page**	
Word	Atten.	Word	Atten.
(Dec)ember	.019	(V)atican	.019
(Feb)ruary	.019	editor	.017
point	.019	(K)orean	.017
(gover)nments	.018	(Be)ijing	.017

Goal 2: Performance on $C^{''}$ For our second task, an expert annotator[7] rates 100 documents from each unlabeled corpora in blind trials with a simple rating $\in \{0, 1\}$ for whether they would assign a journalist to investigate a story based on the document. We report these results in Figure 2. The RT model still outperforms for the Bills corpora (.88 AUC) and the Court-Cases corpora (.7 AUC), which according to Figure 1 are more similar to the training data. The BT model is the top performer for the City Council corpora (.79 AUC). We explore possible explanations in Section **??**.

To gain insight into the attention given by BT, we show in Table 3a the top average attention scores, A_{BT}, across heads in the last layer for all documents in C'-test. A_{BT} where $y = 1$ have emotional salience, like "threatens". In contrast, A_{BT} where $y = 0$ are mainly local words. Local news is often published in the Metro Section: BT learns to distinguish emotional salience and locality.

We run a similar experiment on RT, shown in Table 4a. A_{RT} where $y = 1$ are time-based words. This is another form of newsworthiness: when events are time-stamped, they are more likely to be important in the short-term. A_{RT} where $y = 0$ are international signifiers. The dynamics are parallel to BT: international news is often published in the International section.

Finally, we show coefficients for LR, which are all more topic-based. Top positive β^{+} are political events, while top β^{-} are business terms.[8]

Overall, it appears our models are each learning different aspects of newsworthiness. While RT outperformed in almost all categories of evaluation, future work might point towards different kinds of newsworthiness that can be serviced by different approaches.

[7]Our annotator worked for 4 years as a journalist at a major national newspaper and was involved in page-layout decisions.
[8]The phrase "survived wife" is often used in Obituaries.

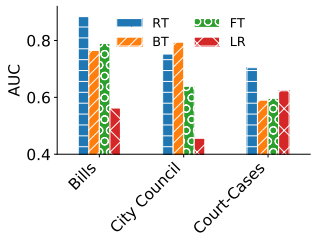

Fig. 2. AUC measured for expert annotation on the external corpora gathered.

Top Pos. Coef.		Top Neg. Coef.	
Word	β	Word	β
nation largest	.25	share earns	-.41
people killed	.25	survived wife	-.41
communist party	.23	media business	-.37
court ruled	.23	share	-.37

Table 6. N-grams receiving the top positive and negative coefficients in the Logistic Regression model.

6 CONCLUSION

In this work, we have formalized a novel classification task, “newsworthiness ranking”, for which ample training data exists. We have translated what was classically a human judgement on the democratic importance of information into an observable metric and modeled it with high accuracy.

We have shed light on some of the factors that contribute to historical judgements on newsworthiness. Such exploration, we observe, has the potential to contribute positively to our information economy by helping both readers and journalists find and consume more socially relevant information.

REFERENCES

[1] Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress toward “the holy grail”: The continued quest to automate fact-checking. In *Computation+ Journalism Symposium, Evanston*.

[2] Sarah Cohen, James T Hamilton, and Fred Turner. 2011. Computational journalism. *Communications of the ACM*, 54(10):66–71.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[4] Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 115–122. IEEE.

[5] Jens Finnas. The hard parts about automating journalism. *Google Slides*.

[6] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.