NewsFuse: Personalized Ranking System of News Stories For Journalists In An Age Of Internet Noise

Aparna Ghosh
Seattle, Washington
ag3294@columbia.edu

Madhura Raju
Seattle, Washington
madraju@microsoft.com

ABSTRACT

News consumers want to be in touch with the happenings of the world at all given times of the day. The old models that news aggregators used to filter and curate news from various sources and report only once a day, are passe now. This changing landscape of news consumption poses two serious challenges for journalists and news organizations: firstly, to be constantly looking for newsworthy stories to provide to their easily-bored readers; and secondly, to be able to cut through the vast volumes of noise -fake and real- on the Internet. To assist in this process, we introduce NewsFuse, an application that suggests and ranks various news stories within the journalist's areas of interest and news source preferences to help decide on what story to work on for the day. This system leverages certain hand-picked NLP features from news articles combined with an Interest classifier, to build a relevance ranking system based on user preferences. We also propose concrete opportunities for the further development of this application and different use cases that it can be used in.

KEYWORDS

Journalism, Beat Reporting, Automated System, Relevant News, Scoring Algorithms, Article Categorization

1 INTRODUCTION

Journalists usually have specific areas of interest called beats that they track closely. Experts and academicians in the past have deemed this term in various ways. While some of them call it just beat reporting, others refer to it as specialized reporting. This kind of reporting of specific topics still forms an integral part of today's journalism. It is a well-known fact (within the journalism community) that the best beat reporters have in-depth knowledge about the issues, people, organizations and trends in their beats.

Some organizations let their journalists track a broader class of beats such as science and technology, business, politics or sports, while some organizations might assign journalists with specific beats such as healthcare in Europe, technology startups, current political affairs or climate change. Irrespective of which category the journalist falls under, either a broad beat reporter or a hyper-specific one, keeping track of the subject and producing relevant content timely is his or her main goal. Today's journalists use Google searches with keywords they have customized (after a lot of trial and error) to read relevant stories on their curated set of news websites.

The dramatic growth of social media platforms let journalists and readers to access news from many different channels as and when new stories break across the globe. This poses a potential risk of journalists getting distracted in the Internet news noise with irrelevant or loosely relevant news stories.

To work on this issue, we propose NewsFuse, an application to enable journalists to see rankings of what the popular or important stories of the day in their beats or interest areas are, and thus create a customized tracking experience. To further curate the list of stories, journalists will first provide inputs of the names of people, organizations or institutes that the journalist is tracking, and the list of all the places (or sources) where the journalists gather their everyday news from. As part of NewsFuse, we have built two different algorithms: a document classification algorithm to categorize the body text of the article; and a scoring/ranking algorithm to score and return the most relevant news titles for the journalist to follow on that day.

To the best of our knowledge, this is one of the first free tools available to use a deep learning approach for article classification. We have conducted multiple sets of tests for the evaluation of the system using journalists' manual judgments to choose their most relevant news stories for the day. Initial results show great potential of this system to support the journalists in the future.

The work in this paper is organized as follows. In Section 2, we discuss relevant recent work on automated journalism support, document classification, and existing systems in the market that perform similar functions. We then present our system with information on its overall architecture, including details on algorithms used and user experience in Section 3. Section 4 presents the manual testing and evaluations of our system. Section 5 summarizes the scope of future work, and finally Section 6 has the conclusion.

2 NEWS GATHERING TECHNIQUES

To help us better understand the news gathering techniques journalists used everyday, we performed interviews with 15 different (nine female and six male) beat reporters including science and technology reporters, business reporters, arts and culture reporters and politics reporters. Out of the 15 journalists, seven are graduate students with some prior experience (at least 3 years) in news and feature reporting, while the other eight are professional journalists either freelancing in a particular beat, or working for various news publications in the USA.

We also used this opportunity to make the journalists test the NewsFuse system.

C+J Symposium, 2017,

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

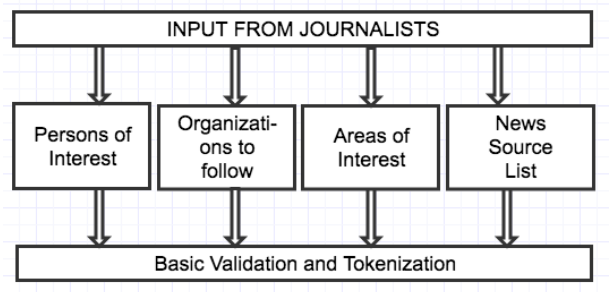

Figure 1: User Experience Block Diagram

3 THE NEWSFUSE SYSTEM

NewsFuse is an online application that provides journalists with a ranking of the most relevant stories within their network and interests, and the frequency of the stories on chosen news platforms. In this section, we will describe the overall user experience and system architecture of the NewsFuse system.

3.1 User Experience

Most journalists today maintain an Excel sheet with the all information of the people they follow or organizations they track. Let us for our convenience call it the MIS (master information sheet). When the journalist registers on NewsFuse, they are asked to either link their MIS to the NewsFuse system or customize their interests. The journalist is asked to build a list of a maximum of 15 different news websites (includes competitor platforms) they source story ideas from daily. With the information on the MIS and the list of source websites, NewsFuse runs search algorithms on the websites to find elements of interest. The search algorithms return the top 5 high frequency results (with the number of occurrences of similar stories in the selected source websites so far) on the Stories Of The Day list. The NewsFuse system will also send out a brief of the results to the journalists email every morning at a time decided by the journalist. The Stories Of The Day list is updated real time with every refresh.

3.2 System Architecture

For the sake of a prototype we will limit the number of 'elements of interest' field to 10 (five persons and five organizations), and the various 'sources' to 20. The application uses the elements of interest to perform a search from the text gathered from various news sources. A free API platform (newsapi.org) provides access to about 70 different news websites' APIs to help gather text from the list of sources. We are continuously on the lookout for more such news API aggregators, to be able to add more publications' websites for the journalist to choose from.

With the proposed NewsFuse, the process of generating a curated list of stories for the journalist takes place in three stages. The first stage of the system is the article or text classification, where the system searches the publication websites for articles published in the last 24 hours, and gets an output with the entire text of the article, to decide the categories of the articles. In this case, it puts the articles under one of the categories: business, entertainment, politics, sport, or technology. The second part involves the entity

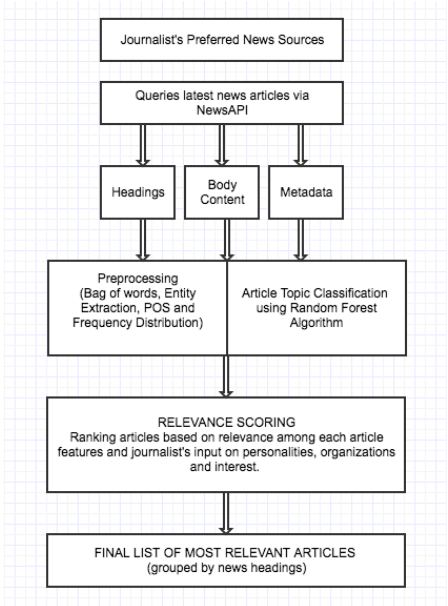

Figure 2: System Architecture Block Diagram

matching, where nouns of names and organizations are extracted from the text of the articles. We will then correlate the extracted entities (names and organizations, in this case) with the elements of interest input by the journalist. Here we use the Bag-of-Words, a set of pre-processing like lemmatization, POS tagging, stop word removal and Named Entity Recognition to check for text similarity, and return a raw score for every relevant URL with the journalist's input query. The last stage is grouping the top relevant articles obtained from the previous step by topic, so the journalist would have a list of topics to start with at every instance respecting their source list and interests.

This will enable journalists to decide what the most important story of the day is, and what he or she should be working on or following up on for the rest of the day.

3.3 Algorithms

Step 1: Article Classification

To accommodate the field of interests that the journalist follows, we built a quick topic classifier that would given an article identify the topic. To train the classifier, we employed the BBC dataset that focuses on five topics - Business, Politics, Sports, Technology and Entertainment. The dataset contained approximately 400 articles for each of the topics. For Feature selection that is used as an input to the modeling after analyzing the dataset we zeroed down on the following features: tokens with stop words removed, top 100 frequently mentioned bigrams, using Stanford's Named Entity Recognition we picked the tokens that matched Person, Location and Organization and Term Frequency Inverse Document Frequency as our final features to the input algorithm.

We used a simple Random Forest algorithm for training the data and used a validation set of 30 articles from each topic to tune the random forest parameters like tree depth and number of trees. Testing the classifier on test set that contained around 50 articles

Using the interests specified in the input query by the journalist, we generated a cluster of related entities using synonyms, nouns and adjectives extraction from the definition of the words. If the lemma of this cluster contained the interest tag from the specific article, then this article was scored higher than the other candidates.

Step 2: Article Scoring

This forms the crux of the relevant ranking algorithm. When extracting the news articles, we pay specific attention on title and the article body itself. Using NLTK, we tokenize the words followed by a set of preprocessing steps. Some of these include: removing stop words that might over value most frequent words, lemmatization based on the Part of Speech to get the word lemmas, removing punctuations, identifying Part of Speech Tags and using Entity Recognition algorithms to obtain Person and Organizations in each of the article lists we gathered. Once the preprocessing steps were complete, we scored each article based on the intersection of tokens, bigrams and NER matchings between the article and the journalist's initial inputs. If these bigrams were present in the headings, we gave it more weight. We performed these steps for both person and organizations and queried only on the article of 'interests' based on the article classifier.

Step 3: Clustering by Topic

With the list of relevantly ranked articles for every source based on the input parameters, we finally group the articles based on the title topics. Using Cosine Similarity between the titles and Entities of every article, we group the articles into buckets of different topics. For every article, we calculate the similarity metrics for every other relevant article and understand the currently trending topics. The journalist would receive the topics with the list of URLS that would help get jumped started. Comparing with Jaccard and Dice similarity, Cosine seemed to perform well for the samples we experimented with.

4 EVALUATION

4.1 Golden Standard: Picking The Most Relevant Stories Of The Day

To evaluate if the suggestions of the NewsFuse system were relevant, we asked the 15 journalists for their inputs, including people of interest, organizations or institutes they tracked, and their beats or categories. The system generated the top five stories for all the journalists. We informed the journalists to acknowledge if the top five stories generated by the system were relevant to them (and if they would work on that story for the day) or not. They could label the story suggestion 1 if it was relevant, or 0 otherwise. We generated 75 stories (5 stories for each of 15 journalists) for different inputs, and evaluated the results of the reliability survey from the journalists.

```
# Journalist's preference
Persons_of_interest = ["Paul Graham", "Elon Musk", "Bill Gates", "Mark Cuban"]
Organizations = ["Microsoft", "Tiger Capital", "Y combinator", "Tesla"]
Source_list = ["techcrunch", "mashable", "the-wall-street-journal", "the verge", "bloomberg", "hacker-news", "recode"]
Interests = ["Technology"]
```

Figure 3: Inputs From Journalist 1

```
Topic: Tesla's first sleek Model 3s are on the road
Related Links:
https://www.theverge.com/2017/7/30/16061540/elon-musk-model-3-names-joke
https://www.recode.net/2017/7/29/16061320/elon-musk-tesla-live-event-model-3-mass-market-delivery
http://mashable.com/2017/07/29/tesla-model-3-first-look/
https://www.bloomberg.com/news/articles/2017-07-31/driving-tesla-s-model-3-changes-everything
https://www.theverge.com/2017/7/29/16060790/tesla-model-3-interior-controls-first-drive-2017
https://www.bloomberg.com/news/articles/2017-07-31/japan-attempts-first-rocket-launch-to-join-spacex'

Topic: Uber's CEO search is down to only male candidates — as its board struggles and Travis Kalanick meddles
Related Links:
https://www.recode.net/2017/7/30/16066332/uber-ceo-search-travis-kalanick-meg-whitman-steve-jobs-board
https://www.recode.net/2017/7/28/16059386/cruise-charlie-miller-chris-valasek-uber-didi
https://www.recode.net/2017/7/29/16059140/stitch-fix-delivery-clothing-ipo-filing-ceo-katrina-lake

Topic: Apple issues statement regarding removal of unlicensed VPN apps in China
Related Links:
https://techcrunch.com/2017/07/30/apple-issues-statement-regarding-removal-of-unlicensed-vpn-apps-in-china/
https://techcrunch.com/2017/07/29/apple-removes-vpn-apps-from-the-app-store-in-china/
https://www.theverge.com/circuitbreaker/2017/7/30/16067172/next-iphone-screen-design-face-unlock-confirmed-homepod-firmware
https://www.wsj.com/articles/iphones-toughest-rival-in-china-is-wechat-a-messaging-app-1501412406
https://www.theverge.com/2017/7/30/16058174/diy-cellphone-sniffer-gsm-imsi-catcher
```

Figure 4: Top 3 Stories With URLs For Journalist 1

```
# Input: Journalist's preference
Persons_of_interest = ["Donald Trump", "Vladimir Putin", "Emmanuel Macaron", "Pope Francis", "Narendra Modi"]
Organizations = ["United Nations", "Supreme Court", "Republican Party", "Democratic Party", "Kennedy School"]
Source_list = ["abc-news-au", "bbc-news", "associated-press", "reuters",
               "the-guardian-uk", "the-washington-post", "usa-today", "daily-mail"]
Interests = ["politics"]
```

Figure 5: Inputs From Journalist 6

```
Topic: Donald Trump is wielding the knife and governing with abandon, despite chaos in the White House
Related Links:
http://www.abc.net.au/news/2017-08-07/donald-trump-is-wielding-the-knife-and-governing-with-abandon/8780662
https://www.reuters.com/article/us-usa-trump-pence-idUSKBN1AM0N3
https://www.washingtonpost.com/news/politics/wp/2017/08/06/flake-extends-his-attacks-on-trump-and-the-gop-all-the-way-back-to-the-dawn-of-birtherism/
https://www.washingtonpost.com/news/fact-checker/wp/2017/08/07/president-trumps-claim-of-obamacare-bailouts-for-insurance-companies/
https://apnews.com/76dd85a360a243950a30a13d71237b02

Topic: Governor calls mosque bombing  'act of terrorism'
Related Links:
http://www.sctimes.com/story/news/local/2017/08/06/governor-calls-mosque-bombing-act-terrorism/543600001
https://apnews.com/6302c6574019dfe6b055935bca59f3331
https://www.reuters.com/article/us-northkorea-missiles-media-idUSKBN1AN03N

Topic: China media stress limits to North Korea sanctions, slam U.S. 'arrogance'
Related Links:
https://www.reuters.com/article/us-northkorea-missiles-media-idUSKBN1AN03N
https://www.washingtonpost.com/world/national-security/china-urges-north-korea-to-be-smart-and-drop-its-missile-tests/2017/08/06/fa2e90c4-3bff-4877-a190-6acaa2acd179_story.html
https://www.reuters.com/article/us-northkorea-missiles-idUSKBN1AN054
https://www.theguardian.com/world/2017/aug/07/we-are-powerless-un-syria-investigator-carla-del-ponte-quits-over-lack-of-political-backing
```

Figure 6: Top 3 Stories With URLs For Journalist 6

4.2 Implementation

We present in this section two separate test cases of the implementations of input retrieval from journalists. We received inputs including the people of interest, organizations or institutions that they track and sources or news websites to track the stories on from a set of 15 journalists.

Figure 3 and 4 represent the input and outputs of the test case for Journalist 1, a technology reporter. Figure 5 and 6 represent the input and outputs of the test case for Journalist 6, a politics reporter.

4.3 Analysis

We performed analysis of the results of the reliability survey from the journalists. Out of the 75 stories suggested by the NewsFuse system, journalists found 59 stories in line with what they would have worked on if they had curated the story list manually. In other words, 59 out of 75 stories were scored with 1s, and 16 stories got 0s. We then calculated the average of the scores for all 15 journalists independently.

Table 1 shows the reliability survey results from the 15 different practicing beat reporters in various organizations. The values ranged from 0.6 to 1 for separate journalists, and had an overall

Table 1: Reliability Score Survey Results

Journalist	Area of interest	Reliability Score
Journalist 1	Technology	0.8
Journalist 2	Business	0.6
Journalist 3	Business	0.8
Journalist 4	Technology	0.8
Journalist 5	Entertainment	1.0
Journalist 6	Politics	0.8
Journalist 7	Sports	0.8
Journalist 8	Sports	0.6
Journalist 9	Entertainment	0.8
Journalist 10	Politics	0.6
Journalist 11	Business	1.0
Journalist 12	Business	0.8
Journalist 13	Technology	0.8
Journalist 14	Technology	0.8
Journalist 15	Politics	0.8
	Average Reliability Score	0.786

average reliability score of 0.786. This means that the system (under the current constraints) produces an accurate personalized rankings of stories approximately 79 percent of the times.

5 FUTURE WORK

To make the NewsFuse more effective in the future, we propose various extensions and additional features to the discussed system above. This system could have an additional feature to introduce journalists to newer related interests, personalities and organizations that might come within the journalist's radar. By finding similarity in the journalist's preferences and the existing trending news, this system would expand the realm of the journalist's domain. These suggestions could be ranked based on the trending news or potentially viral news subjects. This would be an add-on to introduce similar entity recommendations to the journalist.

This system also helps in combining different news articles of the same topic together to give the journalist a grouped view of his daily digest. Though this makes the system credible, there could be cases when falsified news could exist in many of the journalist's choice of newspapers. We would like to build a fact checker system, in between the journalist's input and the final ranking results, which would take the authenticity of the article into account in the Ranking Algorithm.

NewsFuse is in the Beta stage. The article classifier needs to be trained with a dataset that has more granular categories and we plan to explore deep learning or neural network algorithms, so feature engineering is more inbuilt.We plan to explore in more detailed the research around text similarity and topic classification, in order to improve the accuracy of the overall system.

The cosmetic changes to this system would be to build it as a RESTFul API so any web application or a mobile application would connect to it, retrieving relevant images to make it look more appealing to the journalist. They would be able to bookmark articles for future reference, share the list to a co-worker, make notes and like articles. These likes could be used for refining the recommendation systems. For every article, they would also be able to see the twitter feeds based on the hashtags created from the entities on the article.

As automatic summarization has become widely explored field, to make it easier for the journalist we could also summarize the text from different articles on the same topic, while grouping the news articles.

6 CONCLUSION

We have presented NewsFuse, an online tool for a personalized ranking of news stories for journalists based on their preferences. The NewsFuse system using three algorithms shows reasonable performance in its evaluation. From the evaluation, we calculated the average relevance score of the system to be approximately 0.79 for suggested articles. We discussed both the limitations of the system and also suggested various scopes for more advanced technology and future research in the areas of personalized ranking systems for journalists to help with pre-reporting and curation of articles in any given beat and trend recommendations.

REFERENCES

[1]How social media is reshaping news, Pew Research Center, 2014.
[2] Web document classification by keywords using random forests. Myungsook Klassen and Nikhila Paturi
[3] Automating News Content Analysis: An Application to Gender Bias and Readability, JMLR Workshop and conference proceedings.
[4] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.
[5] NLTK: the Natural Language Toolkit. Proceeding ETMTNLP '02 Proceedings of the ACL-02 Workshop in Effective tools and methodologies for teaching natural language processing and computational linguistics