# Finding News Curators in Twitter


Finding News Curators in Twitter ..... 1
1. abstract ..... 3
1.1. main subject ..... 3
1.1.1. Discover news curators ..... 3
1.2. news curators ..... 3
1.2.1. monitor a large variety of sources on a topic or around a story ..... 3
1.2.2. carefully select interesting material on this topic ..... 3
1.2.3. Share it with interested audience ranging from thousands to millions ..... 3
1.2.4. A famous example ..... 3
1.3. characterisitcs of Twitter curators ..... 3
1.3.1. Topic-focused/unfocused ..... 3
1.3.2. With/without user commentary. ..... 4
1.3.3. Human/automatic ..... 4
1.4. contributions ..... 4
1.4.1. 01-Define some of the roles people play in news crowds ..... 4
1.4.2. 02-study data from two large international news organizations and characterize these roles ..... 4
1.4.3. show that to some extent we can statistically model these roles ..... 4
1.5. steps ..... 4
1.5.1. 01-Introduces a set of user classes, which are studied by hand-coding a sample from the dataset ..... 4
1.5.2. 02- described a dataset used ..... 4
1.5.3. 03- obtain the results ..... 4
1.5.4. 04- use labeled data to build automatic news story curator detectors ..... 5
1.5.5. 05-outlines related works ..... 5
1.5.6. 06-Presents the conclusions ..... 5
2. 4- DATASETS ..... 5
2.1. Source ..... 5
2.1.1. two online news websites ..... 5
2.2. collecting procedure ..... 5
2.2.1. downloaded periodically the homepage of each of these websites, ..... 5
2.2.2. get headline news ..... 5
2.2.3. get sampled at random a subset of news articles. ..... 5
2.2.4. sampled articles ..... 5
2.3. Data Cleaning ..... 6
2.3.1. News crowds extraction ..... 6
2.3.2. Spam filtering. ..... 6
2.3.3. Exclude articles ..... 6
2.3.4. kept articles with ..... 7
2.3.5. users with at least 1,000 followers, ..... 7
2.4. LABELING NEWS STORY CURATORS ..... 7

2.4.2. two articles for dataset ..... 7
2.4.3. Labeling process ..... 8
3. 5 Features ..... 10
3.1. network-based features ..... 10
3.1.1. number of followers of a user ..... 10
3.2. contextual features ..... 10
3.2.1. user prole ..... 10
3.3. visibility ..... 11
3.3.1. number of followers ..... 11
3.3.2. number of Twitter lists containing that user ..... 11
3.4. Tweeting activity ..... 11
3.4.1. number of tweets per day ..... 11
3.4.2. fraction of retweet mark ..... 11
3.4.3. fraction of URL ..... 11
3.4.4. fraction of user mention ..... 11
3.4.5. fraction of hashtag ..... 11
3.5. Topic focus ..... 11
3.5.1. many different articles ..... 11
3.5.2. number of distinct sections of the crowds ..... 12
4. 6- Model ..... 12
4.1. Predict ..... 12
4.1.1. UserIsInterestedInStory ..... 12
4.1.2. UserIsHuman ..... 12
4.2. Model ..... 12
4.2.1. Single input feature ..... 12
4.2.2. all input features ..... 13
5. Future work ..... 13
5.1. other variables can be incorporated ..... 13
5.2. a user-centered method could be developed in which the input is a user and the output is a set of news curators. ..... 13
5.3. Check automatic curators functionality in Twitter ..... 13

# 1. abstract

## 1.1. main subject

### 1.1.1. Discover news curators

## 1.2. news curators

### 1.2.1. monitor a large variety of sources on a topic or around a story

### 1.2.2. carefully select interesting material on this topic

### 1.2.3. Share it with interested audience ranging from thousands to millions

### 1.2.4. A famous example

**Andy Carvin (@acarvin),**

**collects news related to the Arabic world, during the Arab Spring**

**Qatari with 2022**

## 1.3. characterisitcs of Twitter curators

### 1.3.1. Topic-focused/unfocused

**topic-unfocused curator**

**collects contents about diverse topics**

**seminating news articles about breaking news**

**topic-focused curator**

**who collects interesting information with a speci**

**c focus**

**geographic region**

**topic.**

**1.3.2. With/without user commentary.**

**include link to the news**

**provide personal comments and opinions**

**1.3.3. Human/automatic**

**automatic news aggregators.**

**@BreakingNews**

**focused on specific topics,**

**they do not provide any personal contents or opinions.**

**1.4. contributions**

**1.4.1. 01-Define some of the roles people play in news crowds**

**1.4.2. 02-study data from two large international news organizations and characterize these roles**

**1.4.3. show that to some extent we can statistically model these roles**

**1.5. steps**

**1.5.1. 01-Introduces a set of user classes, which are studied by hand-coding a sample from the dataset**

**1.5.2. 02- described a dataset used**

**1.5.3. 03- obtain the results**

#### 1.5.4. 04- use labeled data to build automatic news story curator detectors

#### 1.5.5. 05-outlines related works

#### 1.5.6. 06-Presents the conclusions

## 2. 4- DATASETS

### 2.1. Source

#### 2.1.1. two online news websites

BBC World Service (BBC)

Al Jazeera English (AJE).

### 2.2. collecting procedure

#### 2.2.1. downloaded periodically the homepage of each of these websites,

#### 2.2.2. get headline news

#### 2.2.3. get sampled at random a subset of news articles.

#### 2.2.4. sampled articles

use Twitter's API

**nd tweets containing that article's URL.**

### 2.3. Data Cleaning

#### 2.3.1. News crowds extraction

**News crowds**

**all the users who tweeted its URL within the first 6 hours after publication**

#### 2.3.2. Spam filtering.

**discarded accounts having**

**more than 98 tweets per day**

**more than 90% of retweets**

**more than 92% of tweets containing URLs**

#### 2.3.3. Exclude articles

**did not generate a signi**

**cant response**

**generated an extremely high one**

**2.3.4. kept articles with**

**for BBC**

**50 to150 users**

**for AJE**

**70 to 360 users**

**2.3.5. users with at least 1,000 followers,**

**2.4. LABELING NEWS STORY CURATORS**

**2.4.1.**

Table 2: Example of users for two news articles. We include the number of followers, tweets per day, fraction of tweets containing URLs and user mentions ("@"), the type of tweet generation and the main topic.

| | Foll. | Tweets /day | Fraction URL | @ | Type | Topic |
|---|---|---|---|---|---|---|
| **16 Jan 2013 – Syria allows UN to step up food aid** | | | | | | |
| @RevolutionSyria | 88122 | 189.13 | 0.86 | 0.02 | Auto. | Syria |
| @KenanFreeSyria | 13388 | 9.29 | 0.74 | 0.28 | Human | Syria |
| @UP_food | 703 | 10.22 | 1.00 | 0.00 | Auto. | Food |
| **18 Jan 2013 – US cyclist Lance Armstrong admits to doping** | | | | | | |
| @KevinMcCallum | 15287 | 60.15 | 0.18 | 0.77 | Human | Sports |
| @huyanxing | 3224 | 69.19 | 1.00 | 0.00 | Auto. | Misc. |
| @WaseemMansour | 1298 | 15.33 | 1.00 | 0.00 | Auto. | Misc. |

**2.4.2. two articles for dataset**

**civil war in Syria**

**@RevolutionSyria**

**user mentions (\@")**

0,02

**tweets per day,**

189.13

**@UP_food**

**tweets anything containing the word \food"**

**not relevant**

**doping scandal of Lance Armstrong**

**@KevinMcCallum**

**user mentions (\@")**

0.77

**@huyanxing**

**user mentions (\@")**

0

**2.4.3. Labeling process**

**sample of 20 news articles:**

**10 from AJE,**

**10 from BBC**

**for each article**

**sample random 10 users who posted the article**

**tweets that were posted directly after the news article,**

**number of follower**

**profile description**

**three volunteers to provide labels.**

**Q1**

**User interested in the topic ??**

Yes

most tweet related to the topic

May Be

many tweet about the topic

No

not tweeting about the topic expect this tweet

Unknown

can not label

**Q2**

**human ?**

Human

user has conversations and personal comments in his tweets.

Maybe automatic

Automatic

Unknown

**discard may be and unknown**

Table 3: Distributions of the human-provided labels.

| | Interested? | | | Human or Automatic? | | |
|---|---|---|---|---|---|---|
| Dataset | $n$ | yes | not | $n$ | human | automatic |
| AJE | 63 | 21% | 79% | 71 | 55% | 45% |
| BBC | 58 | 3% | 97% | 54 | 35% | 65% |

## 3. 5 Features

### 3.1. network-based features

#### 3.1.1. number of followers of a user

### 3.2. contextual features

#### 3.2.1. user pro

**le**

**3.3. visibility**

**3.3.1. number of followers**

**UserFollowersQ**

**3.3.2. number of Twitter lists containing that user**

**UserListedQ**

**3.4. Tweeting activity**

**3.4.1. number of tweets per day**

**UserTweetsDailyQ**

**3.4.2. fraction of retweet mark**

**UserFracRetweets,**

**3.4.3. fraction of URL**

**UserFracURL,**

**3.4.4. fraction of user mention**

**UserFracMention**

**3.4.5. fraction of hashtag**

**UserFracHashtag**

**3.5. Topic focus**

**3.5.1. many different articles**

**UserCrowdsQ**

#### 3.5.2. number of distinct sections of the crowds

**UserSectionsQ**

## 4. 6- Model

### 4.1. Predict

#### 4.1.1. UserIsInterestedInStory

#### 4.1.2. UserIsHuman

### 4.2. Model

#### 4.2.1. Single input feature

**Simple model**

**UserIsHuman**

$$UserFracURL >= 0.85 \Rightarrow \texttt{automatic}, \text{otherwise } \texttt{human}.$$

**Recal .85**

**Precision 0.85**

**AUC 0.81**

**UserIsInterestedInStory**

$$UserSectionsQ >= 0.9 \Rightarrow \texttt{not-interested}, \text{otherwise } \texttt{interested}$$

**Recal 0.48**

**Precision 0.93**

**AUC 0.83**

**random forest**

**UserIsHuman**

**AUC 0.93**

**4.2.2. all input features**

**random forest**

**UserIsInterestedInStory**

**AUC 0.90**

**4.2.3.**

Table 4: Evaluation of models for the *UserIsHuman* and *UserIsInterestedInStory* tasks.

| | Precision | Recall | AUC |
|---|---|---|---|
| automatic | 0.88 | 0.84 | 0.93 |
| human | 0.82 | 0.86 | 0.93 |
| interested | 0.95 | 0.92 | 0.90 |
| not-interested | 0.53 | 0.67 | 0.90 |

## 5. Future work

**5.1. other variables can be incorporated**

**5.2. a user-centered method could be developed in which the input is a user and the output is a set of news curators.**

**5.3. Check automatic curators functionality in Twitter**