Monitoring trending news and analyzing it's reach to the popular social networking site Reddit

Krupa Ramesh Sawant†
Computer Science
Binghamton University
Binghamton, New York, United States
ksawant1@binghamton.edu

Akshay Kumar Anvekar
Computer Science
Binghamton University
Binghamton, New York, United States
aanveke2@binghamton.edu

Prathamesh Vasant Walke
Computer Science
Binghamton University
Binghamton, New York, United States
pwalke1@binghamton.edu

ABSTRACT

Reddit, also known as the 'Front Page of The Internet', is an online sharing and discussion platform with over 330 million users as of 2018. The site contains a huge collection of text posts, link posts, images, and videos.

Reddit is divided into different topic boards, called 'subreddits'. Each is themed around a different interest or idea. Using a point system of upvotes and downvotes, the community determines which content and discussions are important and subsequently displays them at the top of the feed.

Our project focuses on searching and collecting trending news topics over the internet from a variety of sources like the NY Times, Guardian News and Reddit over a period to measure the amount of user interaction it receives in the form of conversation threads, likes, and shares.

With a decline in traditional television viewing, online social media and news websites like CNN, NY Times have become popular sources of media consumption. As a result, they can be considered as starting points for dissemination of information over social networking platforms and vice versa.

Based on the statistics and analysis of collected data, we will be able to provide insights and useful information that studies user activity, user reaction to the most popular news and the time up to which an article thread may last.

1 Data Sources

We now present the data sources to be used in our measurements.

1.1 The Guardian

The Guardian, formerly The Manchester Guardian, is an influential daily newspaper published in London, generally considered one of the United Kingdom's leading newspapers. Total visits to the Guardian are 262.59M with an average visit duration of 03:43 seconds as of 2019. 10.67% of traffic is from social networking sites and the audience interests vary from technology, electronics, and news.

We are using an API to access all the content the Guardian creates, categorized by tags and section and using a key to successfully authenticate against the API.

1.2 The New York Times

The New York Times is an American newspaper based in New York City with worldwide influence and readership. The New York Times is organized into the following sections: News, New York (metropolitan), Business, Sports of The Times, Arts, Science, Styles, Home, Travel, and other features.
The website had 555 million-page views in March 2005. The domain nytimes.com attracted at least 146 million visitors annually by 2008 according to a Compete.com study. In March 2009, The New York Times website ranked 59th by number of unique visitors, with over 20 million unique visitors, making it the most visited newspaper site with more than twice the number of unique visitors as the next most popular site.

NYT provides its own web API's to fetch trending news from its website. We will be mostly using the Top Stories API provided by NYT.

1.3 The Hindu

The Hindu is an Indian English-language daily newspaper owned by The Hindu Group, headquartered in Chennai. It was started as a weekly in 1878 and became a daily in 1889. It is one of the Indian newspapers of record and the second most circulated English-language newspaper in India. As per the data released by the MRUC, *The Hindu* had an all-India total readership of 62,26,000 in the IRS 2019 and an online newspaper readership of about 5% on an all-India basis.
To gather the data, we will be creating a wrapper as The Hindu provides RSS feeds of its top stories.

1.4 CNN

The **Cable News Network** (**CNN**) is an American cable news television channel. It was founded in 1980 by Ted Turner. CNN provides breaking news alerts and the most talked-about stories through the channel and the online platforms. Total visits to the CNN was 523.03M with an average visit duration of 03:58 seconds as of August 2019. 78.68% audience is from the United States with

a 5.35 % traffic brought by social networking sites like Facebook, Twitter, and Reddit.

We will use the NewsGraph API provided by CNN to collect the data.

1.5 Reddit

Reddit is an American social news aggregation, web content rating, and discussion website. Posts are organized by subject into user-created boards called "subreddits", which cover a variety of topics including news, science, movies, video games, music, books, fitness, food, and image-sharing.

Reddit has more than 330 million monthly active users with nearly 82 billion pages viewed annually. It has more than 138 active communities and receives around 2.8 million comments daily.

The Reddit API will allow us to access the user-submitted and rated stories from reddit.com. It will also provide us with advanced functionality, including user account information and subreddit moderation.

2 Data Collection

The plan is to run a Cron- a software utility tool once a day and collect trending articles from the New York Times, CNN, Guardian and the Hindu.

On an average, we observed that there was a minimum of 10 trending news articles from each channel that appeared as separate threads on Reddit. Therefore, as a rough estimation, we may collect up to 40 threads on Reddit every day which adds up to 300-500 threads, 600-800 subreddits and 1000 comments weekly.

Within a month, we may able to collect about 1500 threads, 1800-2200 subreddits and over 15000 comments.

3 Conclusion

Based on the collected data, we may be able to study and draw some important conclusions about user interactivity, engagement and response to trending news worldwide over the internet.