Content Discovery Advertisements: An Explorative Analysis

Raghavender Rao Jadhav Balaji[1], Andres Baveralle[1], Ameer Al-Nemrat[1], Paolo Falcarin[1]

[1] University of East London, London, United Kingdom
{r.jadhav-balaji, a.baravalle, a.al-nemrat, falcarin}@uel.ac.uk

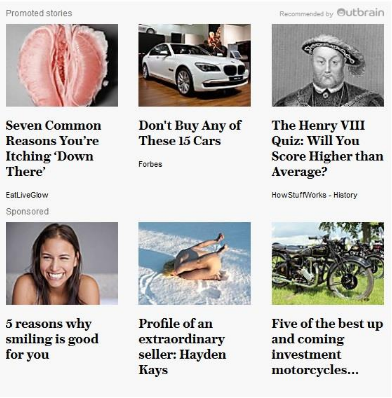
Abstract. Content discovery advertisements are type of native ads which have gained traction for driving ad traffic. These advertisements are being hosted on supposedly reputed websites and their popularity has been growing however it has been reported in the media that these ads are deploying click bait ads. In this research, these ads were evaluated for a period of one month to study and examine their credibility. It was found that significant percentage of these ads were malicious in nature.

1 Introduction

Online advertising has been the one of the significant factors behind the rise of internet marketing and promotion. Online advertisements range from banner ads, widgets, pop-up ads, flash based ads, content ads etc.., One of the driving force of the online advertisements it's easy to place and their reach to millions of users with minimal cost. Online ads have grown in billions in numbers and with ad syndications it has become a complex task to monitor the authenticity of these ads where hackers and malicious persons are able to place deceiving ads on supposedly reputed websites to lure users for commercial gains [1]. Malicious advertising has been a challenge to tackle for internet marketing companies, as malicious ad creators have been able to evolve and evade detection, victimizing millions of gullible users. In this paper we look into content discovery Ads also referred as content recommendation platforms. These advertisements fall under the category of native advertising which have gained more traction and seen on majority of the news and content publishing websites.

1.1 Native Advertising

Native advertising refers to displaying ads that look native to the website and integrate with the content of the webpage. Native ads include sponsored content, branded content, content marketing, and related content [2]. These native ads tend to be more desirable because of the way they are displayed which makes them attracted to companies and users. In 2013 it was noted that 75% of the Online Publishers Association used Native ads [2]. Content recommendation/discovery platforms are new means of advertisements on the webpages. These advertisements are tailored to the contents of the webpages and they blend with the contents of the webpages.



Promoted stories

Recommended by Outbrain

Seven Common Reasons You're Itching 'Down There'

EatLiveGlow

Don't Buy Any of These 15 Cars

Forbes

The Henry VIII Quiz: Will You Score Higher than Average?

HowStuffWorks - History

Sponsored

5 reasons why smiling is good for you

Profile of an extraordinary seller: Hayden Kays

Five of the best up and coming investment motorcycles...

Fig. 1. Example of a Content Discovery Advertisement (Source: www.telegraph.co.uk)

1.2 Content discovery Ads

These ads are commonly found on news and major publishers' websites. They are, usually, wrapped around the contents of the webpages which makes them look as a natural part of the webpage. Figure 1 is a screen shot taken from the popular news website the Telegraph. It showed the content discovery ads on the webpage. The top three articles are content discovery ads while the three articles in the bottom are the news stories from the website itself. These ads look like contents from the Telegraph's website itself but they are, in fact, ads promoted by the content discovery platform providers. These ads can be typically recognized by looking at the URLs of the destination pages (and sometimes also by a small print wrapped around them stating phrases like "from the web" or "promoted stories" and the provider name). The origin of these ads is not always evident to the non-technical users. Content discovery platforms have gained popularity and are often displayed on the supposedly reputed websites. These ads now, as mentioned above, could be seen on many news websites and other content publishing websites. Table 1, illustrates some of these newspapers consumed online that display content discovery ads.

Table 1. Top News Papers Consumed Online in the UK **(**Source: www.pressgazette.co.uk)

News Websites	Users
Daily Mail	10.7 Million
The Guardian	10.5 Million
The Telegraph	9.4 Million
Independent	5.5 Million
Mirror	5 Million
The Express	2.5 Million
Metro	2.5 Million
Huffington Post	1.2 Million
The Sun	0.87 Million

Currently the two major content providers are “Taboola” and “Outbrain”. There are also many other content providers however with a smaller market share [3]. Taboola and Outbrain dominate the content discovery platforms, as of May 2016, with Taboola on at least 24,860 websites and Outbrain on at least 17,746 websites [4] and content discovery ads (CDA) are mostly placed in the category of websites like News, Entertainment, Sports, Technology, business, Media and Forums with ads placed mostly in US followed by the UK [5].

Taboola has become one of the fastest growing content discovery platform company where it reached 1.1 billion clicks in a month and 1 billion unique desktop users [6] [7] and followed by Outbrain, draws clicks from 557 million unique desktops users [8],

1.3 Content discovery platform working model

The content discovery platforms enable small content publishing websites or start up websites to boost their traffic by placing their content as CDA on supposedly reputed news websites such as The Daily Mail, The Daily Telegraph, The Guardian, Forbes, The Sun, CNN, Daily Express, and other similar websites.

One of the boosting feature of these platforms is that the content ads placed on the supposedly reputed websites is tailored to match the relevance of the webpage. Wordings as “related story”, “sponsored story”, or “promoted story” is typically used and may induce the user to click on that particular ad. At the same time, as the ads are included in large traffic, reputable websites, this offers a convenient platform for small or start up content publishing websites to boost their traffic.

These small websites need to pay a fee to the content discovery platforms to place their content. For instance, Taboola operates Cost per click (CPC) model where Taboola charges $ 0.50 per click [9] .

1.4 Credibility of the content ads

As noted earlier in this paper about the growing trend of these ads, several claims have been made by websites and news outlets as the BBC, The Washington Post, The Verge, Fortune, who have reported that these ad platforms are serving malicious advertisements, as click bait scams [10] , and generally low quality content.

Initially CDA were blocked by adblockers as AdblockPlus, Adblock and uBlock. Adblock Plus, one of the most popular adblockers with 100 million active users is now including Taboola’s ads in their acceptable ad whitelist [11].

In this paper, we carry out a critical evaluation of content discovery platforms and their ads on four supposedly reputed news websites from UK. We conducted our analysis in month of April 2016.

2 Methodology

In order to conduct an investigations of the content ads, as mentioned above, four websites that display content discovery platforms, namely: Dailymail, The Daily Telegraph, Daily Express, and The Sun were chosen.

These news websites were chosen for their reputation and popularity, as reflected in their Alexa ranking in April 2016. For our study, we chose top two tiered and bottom two-tiered news website according to the data from Table 1.

Table 2. Alexa Ranking (UK)

Website	*% of Users*	*Ranking*
Dailymail	19.9	19
The Telegraph	27.6	40
Daily Express	36.7	110
The Sun	28.9	156

Taboola ads were displayed on Dailymail website and Outbrain ads were displayed on The Telegraph, Daily Express and The Sun. After identifying the websites to evaluate, we identified content ads on these websites. On each article, Dailymail hosted up to 10 content ads, The Telegraph hosted 3 content ads, Daily Express hosted 6 content ads and The Sun hosted 5 content ads. We examined each ad on all the websites (see Fig 2) and followed the links, and on those links, we identified further content ads, which were sponsored by Taboola, Outbrain and other three companies namely Revcontent, ContentAd and Earnify. On these link pages we identified that content ads were sometimes sponsored by more than one provider.

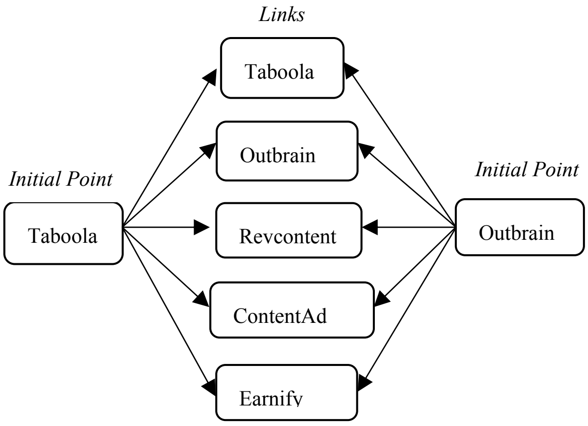


Fig. 2. Content Ad providers on hosting page (initial point) and on Landing page (Links).

2.1 Data Collection:

The first step for the data collection was to scrape the web sites we selected for advertisements. A number of approaches and technologies are described in the literature; from the use of basic string functions and regular expressions to automated extraction of loosely structured data[12].

Due to the dynamic nature of the web sources that we needed to scrape, the use of string functions and regular expressions itself was not sufficient. Instead, JavaScript had to be interpreted and the Selenium library was used for this task. The spider we developed mimicked the user interaction to collect ad data.

We identified the core elements of the main web pages on each website and we identified markers for the spider to locate and capture the advertisement data, using XPath. For instance, the content ad provider Taboola had the following relative XPath: (".//*[contains(@id,'internal_trc_')]") which remained same across all the websites hosting Taboola ads. Once the core markers were identified, we proceeded with our data collection.

Table 3. Examples of the markers.

Content ad providers	Relative XPath
Taboola	(.//*[contains(@id,'internal_trc_')])
Outbrain	(.//*[contains(@id,'outbrain_widget_')
Revcontent	(.//*[contains(@id,'rc_w_')
ContentAd	(.//*[contains(@id,'ac_')
Earnify	(.//*[contains(@id,'earnify-widget-')

2.2 Sequence of data collection:

To run and gather data we set up a Microsoft Windows machine, with no ad block protection. A "real", scripted browser was used for data collection. Basing on[13], the software stack included Python, the Selenium webdriver, Beautiful Soup for data extraction and MySQL to store the results.

The spider did run its sequence for each of the websites under examination every 30 minutes for the period of one month.

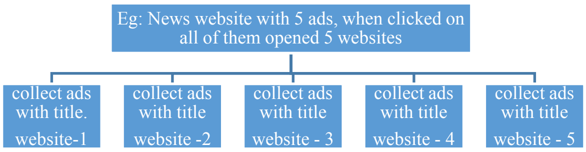

Eg: News website with 5 ads, when clicked on all of them opened 5 websites
collect ads with title. website-1
collect ads with title website -2
collect ads with title website - 3
collect ads with title website - 4
collect ads with title website - 5

Fig. 3. Sequence of data collection

3 Results and analysis:

During the period of our data collection we gathered over hundred thousand content ads. See table 4.

Table 4.

Websites	Total number of Ads
Daily Mail	30057
Daily Express	35219
The Sun	31680
Telegraph	5167

3.1 Taxonomy of the content ads:

In our preliminary analysis we identified that the content ads were displaying different kinds of click bait ads, which we classified with five categories: Money, Sex, Health, Trash and Other.

Money: money scam ads luring users to invest small amount of money in return for high profits in millions.

Sex: sexual content, as fake dating opportunities.

Health: ads luring users to buy products promising to cure variety of body and health related issues.

Trash: ads related to website offering gossip stories; the ad web page hosted further ads related to money, sex, and health.

Other: the ads contained in this category displayed genuine ads.

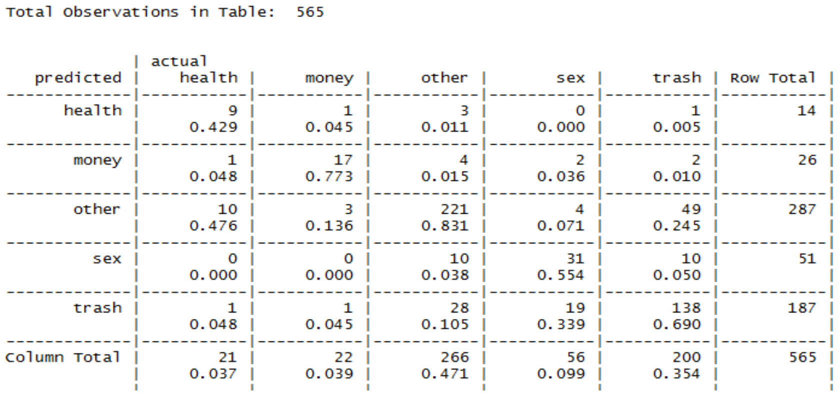
3.2 Classifying the ads:

In our ad classification we used the title of the ads to classify the hundred thousand ads and we implemented Naïve Bayes classifier. In order to train and test the data, we manually classified 1896 out of 30057 ads from Daily Mail, 1708 out of 35219 ads from the Express, 1205 out of 31680 ads from the Sun and 1036 out of 5167 ads from the Telegraph.

3.3 Preparing the data for training and testing:

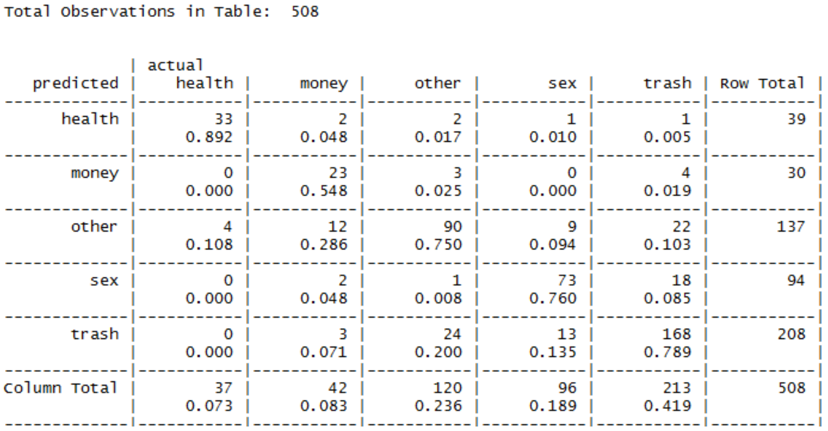
In the Naïve Bayes classifier, we created text corpus for the training data and cleaned the corpus by removing punctuations, numbers and also created file of stop words to be removed from the corpus. Once the corpus was clean we created document text matrix. Then we created training and test data and the Tables 5, 6, 7, 8 shows the results for the respective ads pool.

Table 5. Daily Mail



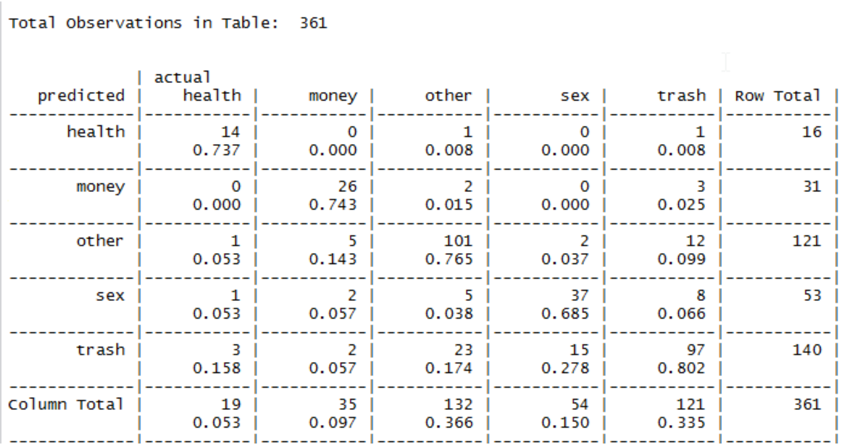
```
Total Observations in Table:  565

             | actual
   predicted |     health |      money |      other |        sex |      trash |  Row Total |
-------------|------------|------------|------------|------------|------------|------------|
      health |          9 |          1 |          3 |          0 |          1 |         14 |
             |      0.429 |      0.045 |      0.011 |      0.000 |      0.005 |            |
-------------|------------|------------|------------|------------|------------|------------|
       money |          1 |         17 |          4 |          2 |          2 |         26 |
             |      0.048 |      0.773 |      0.015 |      0.036 |      0.010 |            |
-------------|------------|------------|------------|------------|------------|------------|
       other |         10 |          3 |        221 |          4 |         49 |        287 |
             |      0.476 |      0.136 |      0.831 |      0.071 |      0.245 |            |
-------------|------------|------------|------------|------------|------------|------------|
         sex |          0 |          0 |         10 |         31 |         10 |         51 |
             |      0.000 |      0.000 |      0.038 |      0.554 |      0.050 |            |
-------------|------------|------------|------------|------------|------------|------------|
       trash |          1 |          1 |         28 |         19 |        138 |        187 |
             |      0.048 |      0.045 |      0.105 |      0.339 |      0.690 |            |
-------------|------------|------------|------------|------------|------------|------------|
Column Total |         21 |         22 |        266 |         56 |        200 |        565 |
             |      0.037 |      0.039 |      0.471 |      0.099 |      0.354 |            |
             |            |            |            |            |            |            |
```

Table 6. Daily Express



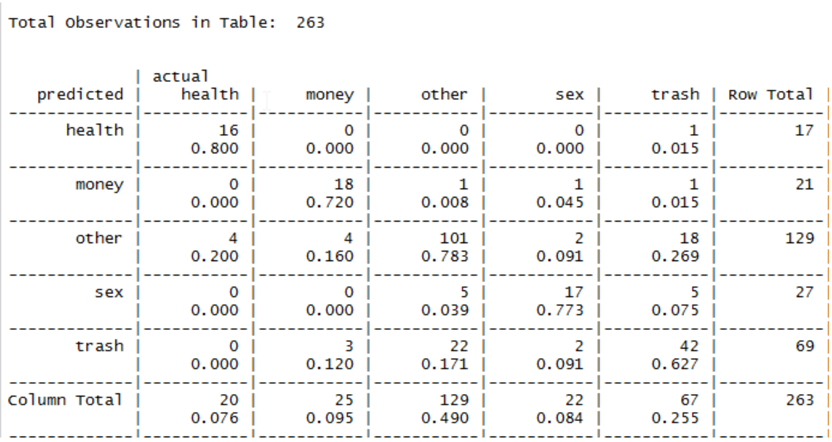
```
Total Observations in Table:  508

             | actual
   predicted |     health |      money |      other |        sex |      trash |  Row Total |
-------------|------------|------------|------------|------------|------------|------------|
      health |         33 |          2 |          2 |          1 |          1 |         39 |
             |      0.892 |      0.048 |      0.017 |      0.010 |      0.005 |            |
-------------|------------|------------|------------|------------|------------|------------|
       money |          0 |         23 |          3 |          0 |          4 |         30 |
             |      0.000 |      0.548 |      0.025 |      0.000 |      0.019 |            |
-------------|------------|------------|------------|------------|------------|------------|
       other |          4 |         12 |         90 |          9 |         22 |        137 |
             |      0.108 |      0.286 |      0.750 |      0.094 |      0.103 |            |
-------------|------------|------------|------------|------------|------------|------------|
         sex |          0 |          2 |          1 |         73 |         18 |         94 |
             |      0.000 |      0.048 |      0.008 |      0.760 |      0.085 |            |
-------------|------------|------------|------------|------------|------------|------------|
       trash |          0 |          3 |         24 |         13 |        168 |        208 |
             |      0.000 |      0.071 |      0.200 |      0.135 |      0.789 |            |
-------------|------------|------------|------------|------------|------------|------------|
Column Total |         37 |         42 |        120 |         96 |        213 |        508 |
             |      0.073 |      0.083 |      0.236 |      0.189 |      0.419 |            |
-------------|------------|------------|------------|------------|------------|------------|
```

Table 7. The Sun



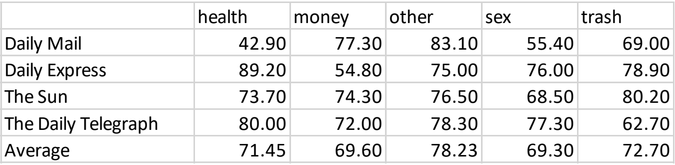
```
Total Observations in Table:  361

             | actual
   predicted |    health |     money |     other |       sex |     trash | Row Total |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
      health |        14 |         0 |         1 |         0 |         1 |        16 |
             |     0.737 |     0.000 |     0.008 |     0.000 |     0.008 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
       money |         0 |        26 |         2 |         0 |         3 |        31 |
             |     0.000 |     0.743 |     0.015 |     0.000 |     0.025 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
       other |         1 |         5 |       101 |         2 |        12 |       121 |
             |     0.053 |     0.143 |     0.765 |     0.037 |     0.099 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
         sex |         1 |         2 |         5 |        37 |         8 |        53 |
             |     0.053 |     0.057 |     0.038 |     0.685 |     0.066 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
       trash |         3 |         2 |        23 |        15 |        97 |       140 |
             |     0.158 |     0.057 |     0.174 |     0.278 |     0.802 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
Column Total |        19 |        35 |       132 |        54 |       121 |       361 |
             |     0.053 |     0.097 |     0.366 |     0.150 |     0.335 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
```

Table 8. The Telegraph



```
Total Observations in Table:  263

             | actual
   predicted |    health |     money |     other |       sex |     trash | Row Total |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
      health |        16 |         0 |         0 |         0 |         1 |        17 |
             |     0.800 |     0.000 |     0.000 |     0.000 |     0.015 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
       money |         0 |        18 |         1 |         1 |         1 |        21 |
             |     0.000 |     0.720 |     0.008 |     0.045 |     0.015 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
       other |         4 |         4 |       101 |         2 |        18 |       129 |
             |     0.200 |     0.160 |     0.783 |     0.091 |     0.269 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
         sex |         0 |         0 |         5 |        17 |         5 |        27 |
             |     0.000 |     0.000 |     0.039 |     0.773 |     0.075 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
       trash |         0 |         3 |        22 |         2 |        42 |        69 |
             |     0.000 |     0.120 |     0.171 |     0.091 |     0.627 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
Column Total |        20 |        25 |       129 |        22 |        67 |       263 |
             |     0.076 |     0.095 |     0.490 |     0.084 |     0.255 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
```

Table 9. Detection rate in %.



	health	money	other	sex	trash
Daily Mail	42.90	77.30	83.10	55.40	69.00
Daily Express	89.20	54.80	75.00	76.00	78.90
The Sun	73.70	74.30	76.50	68.50	80.20
The Daily Telegraph	80.00	72.00	78.30	77.30	62.70
Average	71.45	69.60	78.23	69.30	72.70

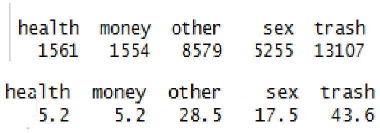
3.4 Classification model:

In our classification model (see table 9) we were able to achieve the detection rate of an average 70% across all categories on all the four websites with an exception in health, sex and money category in the Daily Mail, Daily Express respectively.

3.5 Results

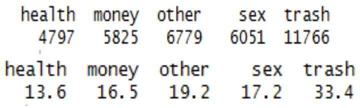
Once the classifier was trained, we applied it to the respective pool of ads we collected, the tables 10, 11, 12, 13 shows the results.

Table 10. Daily Mail in %



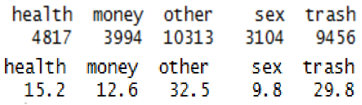
```
health  money  other    sex  trash
  1561   1554   8579   5255  13107

health  money  other    sex  trash
   5.2    5.2   28.5   17.5   43.6
```

Table 11. The Daily Express in %



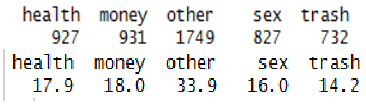
```
health  money  other    sex  trash
  4797   5825   6779   6051  11766
health  money  other    sex  trash
  13.6   16.5   19.2   17.2   33.4
```

Table 12. The Sun in %



```
health  money  other    sex  trash
  4817   3994  10313   3104   9456
health  money  other    sex  trash
  15.2   12.6   32.5    9.8   29.8
```

Table 13. The Telegraph in %



```
health  money  other    sex  trash
   927    931   1749    827    732
health  money  other    sex  trash
  17.9   18.0   33.9   16.0   14.2
```

In the results we identified that once the users clicked on the content ads on the four news websites they were exposed to an average 70% of click bait ads of the category Money, Sex, Health and Trash.

Genuine ads were underrepresented: only 28.5% on the Dailymail, 19.2% on the Daily Express, 32.5% on the Sun 32.5% and 33.9% on the Telegraph.

4 Conclusion:

The content discovery ads, from analysis demonstrated above, are the new gateways for malicious and click-bait websites. In this undergoing work, our aim was to focus on both an exploratory analysis and in the classification of the content discovery ads displayed on the supposedly reputed websites. We also identified that the websites with click bait ads were also hosting malicious ads. Our future work will involve exploring the content ad networks and detecting with precision the malicious ads in this criteria.

5 References:

1. Li, Z., Zhang, K., Xie, Y., Yu, F., Wang, X.: Knowing your enemy: understanding and detecting malicious web advertising. In: Proceedings of the ACM conference on Computer and Communications Security, pp. 674-686. ACM, (2012)

2. Levi, L.: Faustian Pact: Native Advertising and the Future of the Press, A. Ariz. L. Rev. 57, 647 (2015)

3. WSJ, http://www.wsj.com/articles/outbrain-taboola-make-their-mark-on-online-advertising-industry-1426765357

4. Datanyze, https://www.datanyze.com/market-share/content-syndication-networks/outbrain-vs-taboola

5. Builtwith, http://trends.builtwith.com/ads/Outbrain/Market-Share

6. https://www.taboola.com/drive-traffic-and-leads-content-discovery

7. FastCompany, http://www.fastcompany.com/3054205/fast-feed/with-1-billion-monthly-clicks-taboola-is-the-worlds-biggest-content-discovery-tool

8. http://www.outbrain.com/

9. https://www.taboola.com/advertiser-help-center/bidding-and-budget-recommendations

10. BBCNews, http://www.bbc.co.uk/news/business-29322578

11. http://www.ft.com/cms/s/0/80a8ce54-a61d-11e4-9bd3-00144feab7de.html

12. Walther, M.: Unsupervised extraction of product information from semi-structured sources. In: Computational Intelligence and Informatics (CINTI), IEEE 13th International Symposium on, pp. 257-262. (2012)

13. Lawson, R.: Web Scraping with Python. Packt Publishing, Birmingham (2015)