Knowing Me, Knowing User - an analysis of user behaviour on the BBC website during June and July of 2016.

Tomack Gilmore
March 12, 2018

It is fair to say that the summer of 2016 was possibly one of the more eventful of recent years – the UK decided to leave the European Union, Wales nearly made it to the final of the UEFA European Championship, and flash floods damaged two shopping centres in Cheshire. The BBC was, naturally, there throughout this tumultuous time, covering each of these events and more via their website. Data regarding traffic from users through the site during this period is the subject of this report.

Our objective is to build a profile of the kinds of users that visited the BBC website between the 12th of June and the 9th of July 2016. We will do this by analysing a dataset consisting of visits made by a total of 200,000 registered users during this time, focusing on the content they consume, when they consume it, and how. Understanding the data that we have allows us to know our users better, and combining these insights with other machine learning techniques will allow us to deliver to our users more of what they want to see, when they want to see it.

This report is divided into three main sections: in the first we describe the dataset we will work with; in the second we present our analysis and insights into some aspects of the data; and in the third we describe how this analysis could be used to improve services for users and increase their engagement with content on the BBC website. There is also an appendix containing a large number of graphs that were used in our analysis.

1 The Data

The data consists of 15,853,552 records of visits to content on the BBC website made by 200,000 different users between 11pm on the 12th of June and 11pm on the 9th of July 2016. Each record contains the following information:

1. The user id of the person who made the visit;
2. The date and time at which the visit was made;
3. The search term entered into BBC website by the user (if any);
4. The platform on which the content was viewed (mobile, tablet, computer, or big screen);
5. The way in which the content was delivered (web, mobile web, via a responsive app);
6. The kind of content that was viewed (Sport, News, TV and iPlayer, Weather, etc.);
7. The content identifier for the page viewed;
8. The address of the page that was visited;
9. The geographical region where the browser appeared to have arrived from.

There are myriad ways in which we could analyse this data depending on what it is that we would like to try to understand about our audience. As stated above, for the purpose of this report we shall focus purely on content, namely, what types of content are people viewing, when they view it, and how.

2 The Analysis

Before beginning with the analysis proper the dataset needs to be cleaned and tidied, thus it first goes through a pre-processing stage that reduces it down to 13,984,121 visits to content from a total of 192,766 users between 11pm on the 12th of June 2016 and 11pm on the 9th of July 2016. For each visit we will look predominantly at the following features:

1. The user id for each visit.
2. The date and time of each visit.
3. The platform on which the content was viewed.
4. The type of content viewed.

We start by getting a very rough picture of how visits are distributed across different types of content. Figure 2.1 (on page 3) shows the percentage of visits to content in each product category during the entire period, where visits were to one of the following thirteen content categories:

Sport, News, Homepage/search, Tv and iPlayer, Weather, iPlayer Radio, Music, Knowledge, CBBC, BBC Three, About the BBC, CBeebies, Travel, Newsbeat.

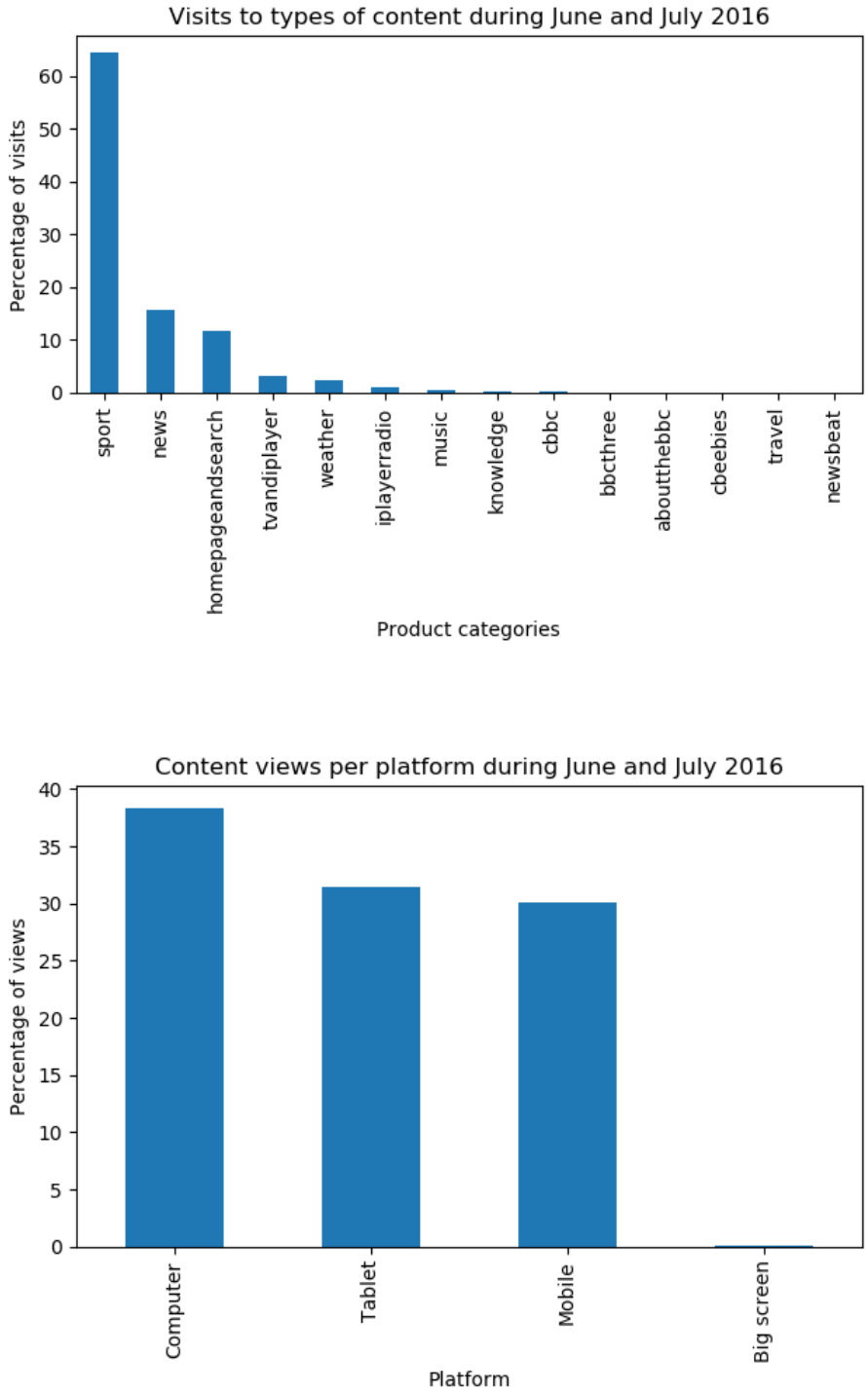

Figure 2.1: The percentage of visits to each product category (above); the percentage of content views per platform (below).

It is clear from looking at the top graph in Figure 2.1 that over 65% of visits to the BBC website in this period were to view sports content, with news coming in at a distant second. This is perhaps not so surprising since both Euro 2016 and the Wimbledon Championship took place during this period.

We could also surmise from the lower graph in Figure 2.1 that very few visits were made on a big screen, in fact content was most often viewed on the computer, and the number of views on mobiles and tablets were roughly the same.

These preliminary charts help to give a general view of what is roughly going on over the whole period, however when we look at little more closely at the data the picture becomes much clearer. Figure 2.2 (on page 5) shows a time series for the entire period, that is, it shows the number of visits per hour to content from one of the top 5 content categories: Sport, News, Homepage and Search, TV and iPlayer, and Weather. We analyse visits to each category below.

Sport: views of sport content peak early in the day, drop off slightly, and then peak much higher later on. The highest volume of traffic was on the 22nd of June, when Hungary drew with Portugal (this matched kicked off at 6pm central European summer time - CEST) and the Republic of Ireland won against Italy (match kicked off at 9pm CEST). The other tallest peaks can be mapped to other matches, for example on the 1st of July Wales beat Belgium in the quarter finals, and on the 6th they lost to Portugal in the semi-finals. Peaks during the week tend to occur more in the evening, whereas on the weekends there are peaks between 2 and 3 and then again in the early evening.

News: during the week (the 13th of June was a Monday) visits to news content appear to be quite regular, it looks as though there is often a mid-morning peak, after which the number of visits decreases slowly through the day. We can see one irregularity towards the end of the 23rd and 24th of June and this is undoubtedly due to the coverage of the EU referendum. On the weekend (18-19, 25-26 June, and 2-3 July) visits to news content are fewer and more or less consistent throughout the day. They are also much in line with visits to the homepage/search page.

Homepage/Searchpage: visits to the home and search pages display similar behaviour to the visits to news content. This is not so surprising, since it is quite likely that people find their way to news content via browsing the BBC homepage or searching for news topics through the search page.

TV and iPlayer: visits to this type of content appear to follow a very regular and consistent pattern, building very gradually across the day and peaking quite sharply in the evening, when users are presumably settling down to watch something. Interestingly this behaviour is consistent across weekdays and weekends.

Weather: it seems that visits to weather content peak in the morning and then decays throughout the day, and much like TV and iPlayer content this seems to hold on weekdays and weekends.

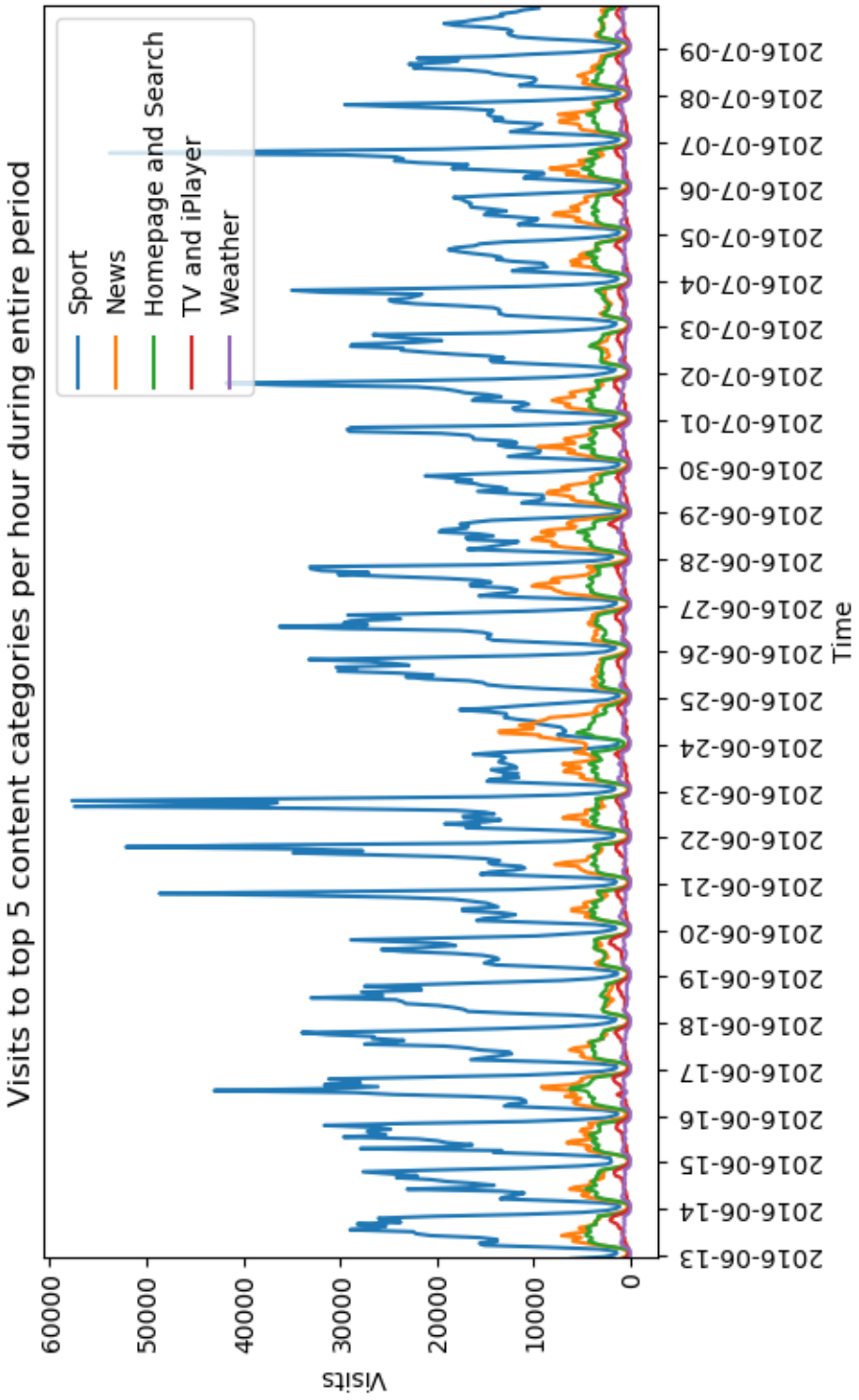

Figure 2.2: Visits to top 5 content categories during June and July.

The behaviour described above can be seen more clearly in the charts in Figure 2.3 (bottom of this page), which show the percentage of visits to each type of content per hour on weekdays and weekends. Most notable are the peaks on weekdays for news content that do not occur on the weekend - between 8 and 9 (presumably people are reading the news before work), then again between 11 and 12 (the mid-morning slump, perhaps?) and then again in the afternoon (a mid-afternoon slump?). We can also see twin peaks for visits to sports content on the weekend that occur around match times (1500 and 2000), while during the week the peak is predominantly during the later evening matches (presumably people are less likely to view sports content at work).

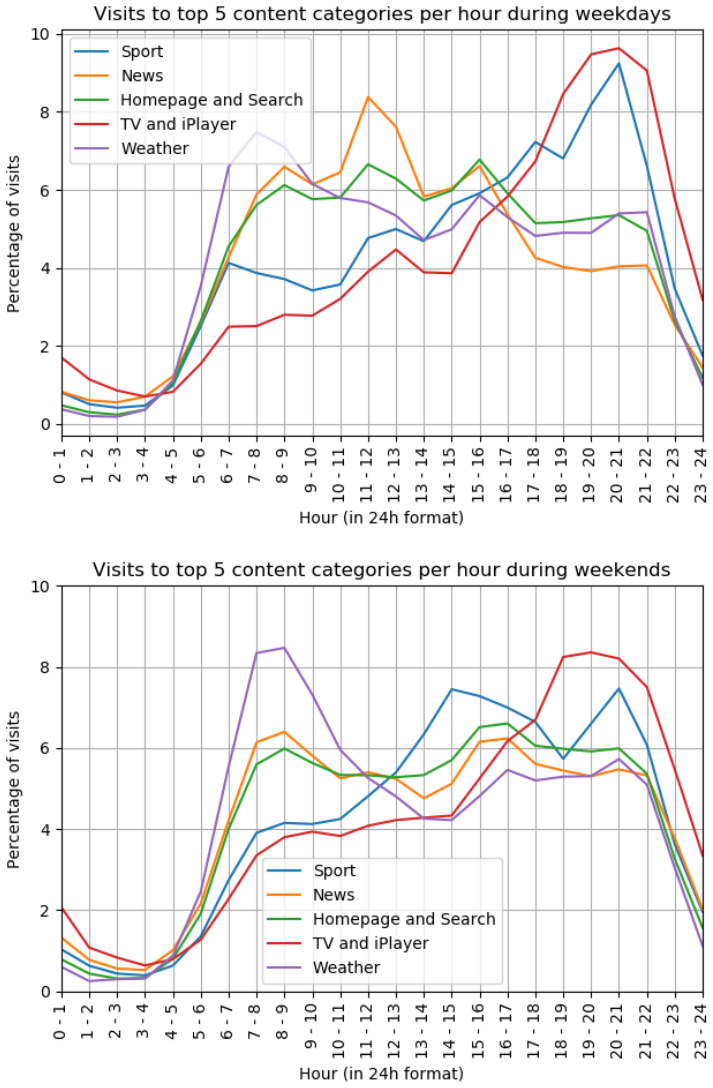

Figure 2.3: Pecentage of visits per hour to each category on weekdays (above) and weekends (below).

2.1 User segmentation

We now have more of an insight into when different types of content are viewed during the week and on weekends, however we know little about how these content views are distributed across users. Our next goal is to try to build up a profile of our users, categorising them according to the types of content they consume.

By using a machine learning technique known as *clustering* we can segment our users based on the proportion of time they spent looking at different types of content during June and July. It turns out that broadly speaking users can be split into four main groups:[1]

- the *sports fans*, who almost always view sports content;
- the *news fans*, who favour news content;
- the *telly fans*, who mainly view TV and iPlayer content;
- and the *homepage searchers*, who spend more time viewing the homepage and searching, checking the weather, and occasionally also viewing sport and news content.

A breakdown of the proportions of our audience that belong to each of these groups is shown in the figure at the bottom of the page (Figure 2.4). Being able to segment our audience in such a way, which comes from the data itself, allows us to really dig deeper into this information in order

[1]These four main groups were selected by using the elbow method – see 'kMeansSSE.png', which was produced using the function 'k_means_cluster_elbow()' from the module 'BBCfunctions.py'.

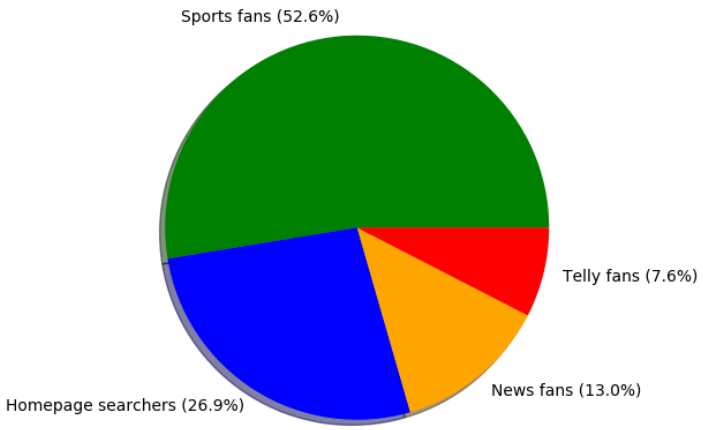

Figure 2.4: Segmentation of users into the four distinct groups (percentages do not quite add up to 100% due to rounding).

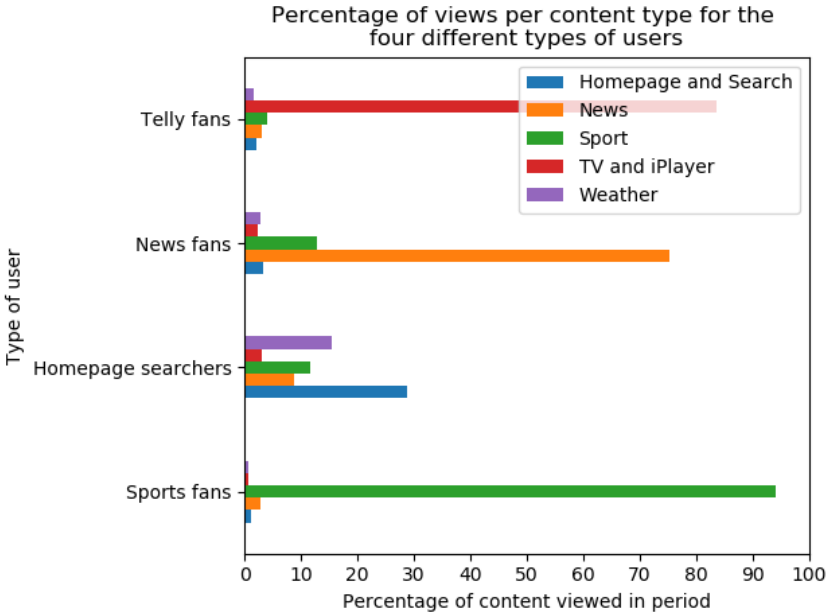

Figure 2.5: Profiling the content views of users from the four groups.

to build up a more accurate picture of how different types of users interact with BBC content.

Roughly speaking the segmentation algorithm looks at the proportions of the types of content viewed by each user over the entire period, and then groups together those users who view similar types of content at similar rates. It is possible to extract from the data the profile of a 'typical' user from each group and a breakdown of how much of each type of content is consumed by a typical member can be seen in Figure 2.5 at the top of this page. A sports fan, for example, typically spends around 95% of her time on sports content, and rarely views anything else from the BBC, while 75% of a news fan's visits are to news content and around 13% are to sports.

Armed with our new ability to profile users we can re-examine content views per hour on weekends and weekdays for users belonging to each group. To do this we have generated quite a large number of graphs that detail the proportion of visits per hour and the platform through which the visits came during weekdays and weekends for each group. There are many of them so we summarise our findings in terms of character profiles for each group below – the graphs themselves can be found in the appendix.

2.2 Profiling each user group

We now offer an overview of how users from each group interact with the BBC website during the course of the day both in the week and on weekends. This follows from studying the figures in the appendix, which were derived by segmenting the data we have into the different groups

of users that emerged from the clustering method discussed above.

The sports fans – these users tend to view mostly sports content through tablets or mobile phones. On a weekday they tend to look at sports content throughout the day, with a peak between 6 and 7am, a gradual decrease until 11-12pm where sports content views pick up again and increase pretty steadily throughout the day until the second (maximal) peak between 8 and 9pm (this was usually the kick-off time for the evening matches during Euro 2016). On weekends there is an increase in views until a peak between 2 and 3pm, a slight decrease until between 6-7pm where the visits increase again, peaking once more between 8 and 9pm. Very few visits come from computers, and there is something of an even split between mobile phone and tablet views throughout the day (see Figure 4.1, 4.2, 4.3, and 4.4 on pages 11 - 14).

The homepage searchers – these users prefer to use computers to view content from the BBC, although they increase their consumption via mobiles and tablets in the evening during the week and on weekends. Another appropriate name for this group could be *workplace browsers* – people who perhaps have the BBC homepage as the homepage of the computer they take to work, and so throughout they day they browse the homepage, search for information, read about news and sport, and consistently check the weather. During weekdays they have a tendency to view more news content in the morning, however as the day wears on this balance shifts and their consumption of sports content overtakes their interest in news after lunch. On the weekend they favour sport over news through the entire day, and the weather is checked at a higher rate between 7 and 9am than in the week (see Figure 4.5, 4.6, 4.7, and 4.8 on pages 15 - 18).

The news fans – these users also favour computers for their BBC content over hand-held devices such as phones and tablets. On weekdays there is a gradual increase in views of news content through the early morning until 8-9am, followed by a slight dip, a maximum peak between 11 and 12pm, and a further peak between 3 and 4pm followed by a slow decrease into the evening. They have a relatively consistent level of engagement with sports content. On the weekend there are more visits made via mobiles and tablets (though the computer is still the main platform), with a peak in the rate of visits to news content again between 8 and 9am, followed by a gradual decrease until 2pm, then an increase and more or less consistent rate from 3pm until 10pm. News fans also have a higher rate of sports content views per hour on the weekend than in the week (see Figure 4.9, 4.10, 4.11, and 4.12 on pages 19 - 22).

The telly fans – the hourly rate of visits to TV and iPlayer content increases gradually from 4am until it hits a peak between 12 and 1pm, then decreases until 3pm after which children come home from school and the rate increases steadily to a maximum peak between 8 and 9pm on a weekday. Telly fans overwhelmingly prefer to use a computer to view content, although mobile and tablet use does increase in the evening. Weekend use is slightly different – the rate of TV and iPlayer content views increases gradually throughout the day until it hits a peak between 7 and 8pm, with a smaller proportion of views being made on mobile phones than in the week. The proportion of visits to this sort of content later into the evening is also higher on the weekend (see Figure 4.13, 4.14, 4.15, and 4.16 on pages 23 - 26).

3 The Potential

Knowing your audience – knowing what content they like to view and when they are more likely to view it – is very powerful indeed and can be utilised in a number of ways. We will focus on two: one approach that primarily benefits the users; and another that is most benefical for stakeholders.

Improving user experience – a common, incredibly successful machine learning technique is that of collaborative filtering, which forms the backbone of modern recommendation systems such as Netflix (and presumably BBC iPlayer). How many times, though, has a system such as this proffered someone an episode of some series, only for them to dismiss it since they were “not in the mood"? By segmenting users in the manner outlined in this report and combining these techniques with a collaborative filter using the original raw data (unfortunately this won’t be expanded upon here, though it is certainly feasible), one could create a recommendation system for users of the BBC website that would recommend content to them at times in the day when they are most likely to view it, offering them more of what they want, when they want it.

Influencing user behaviour – suppose there is some content that is, for whatever reason, not receiving the attention it deserves. By targeting different groups of users at the right time based on their group profile we could maximise the chances of engagement with the new material.

4 Appendix

The following graphs (pages 11-26) were produced using Python and were used to help build a picture of how the different user groups consume content across platforms and content types, at an hourly rate, on weekends and during the week.

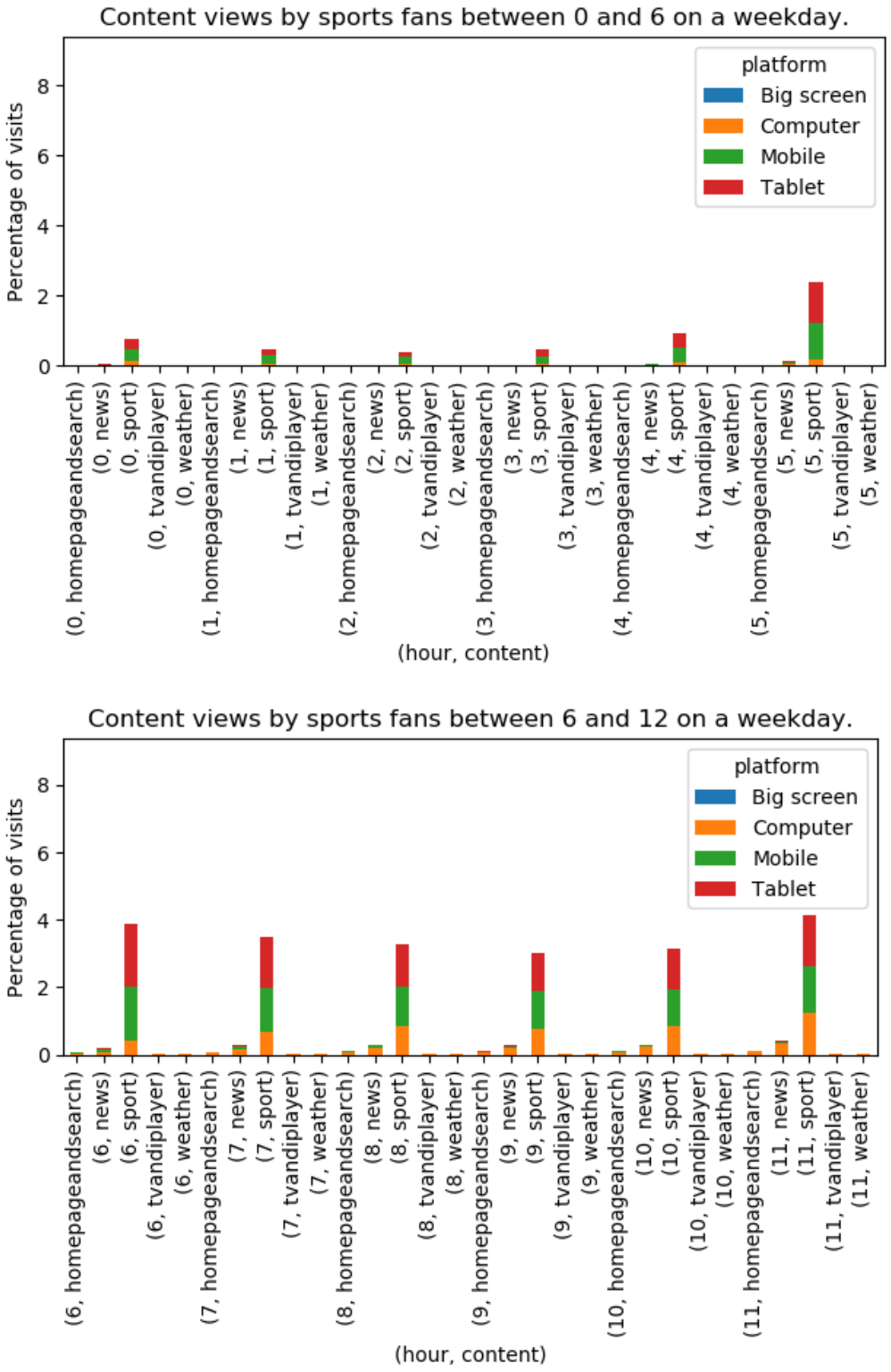

Figure 4.1: Sports fans' content views via platform from midnight to 6am (above) and from 6am to midday (below) on a weekday.

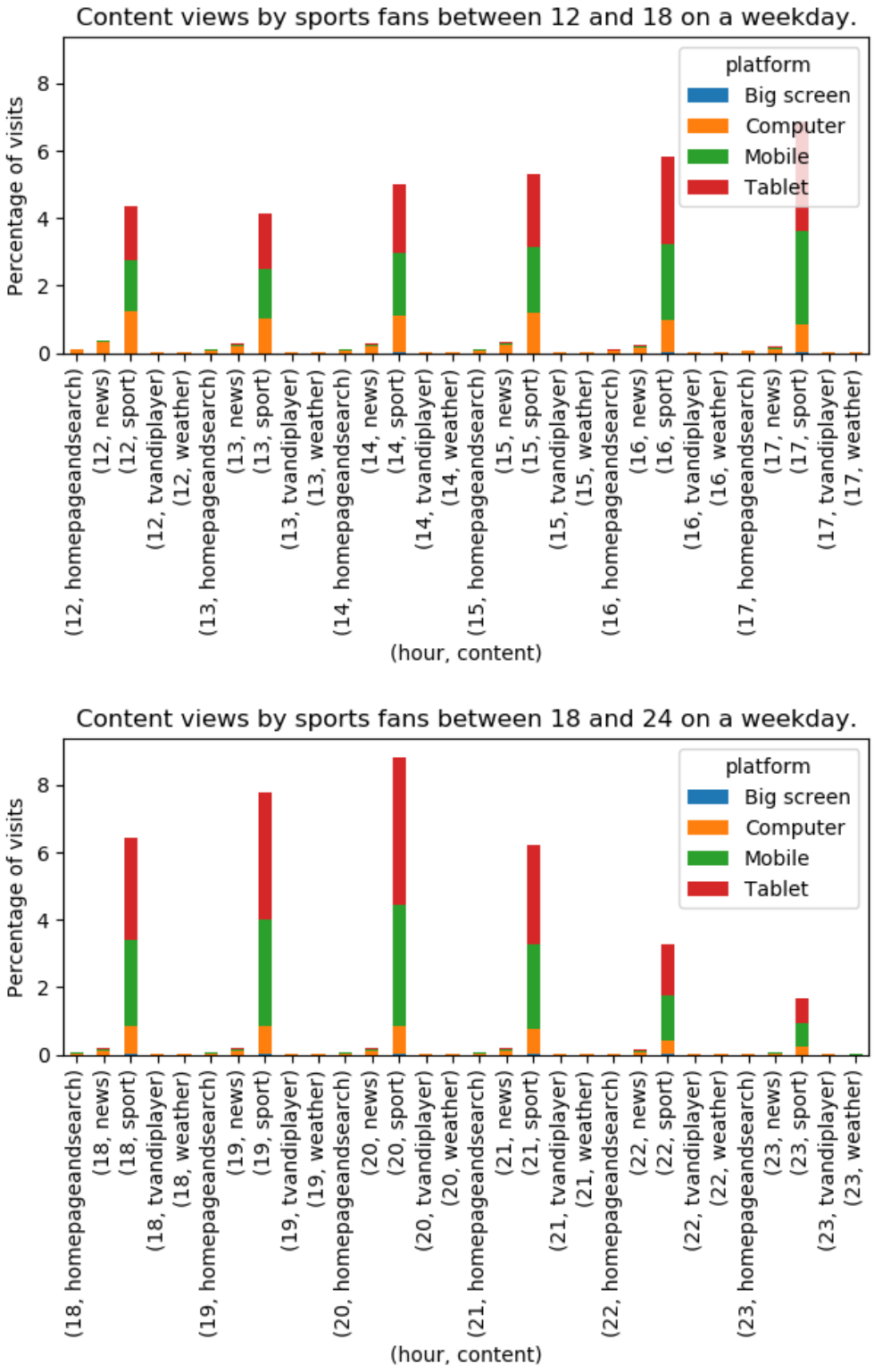

Figure 4.2: Sports fans' content views via platform from midday to 6pm (above) and from 6pm to midnight (below) on a weekday.

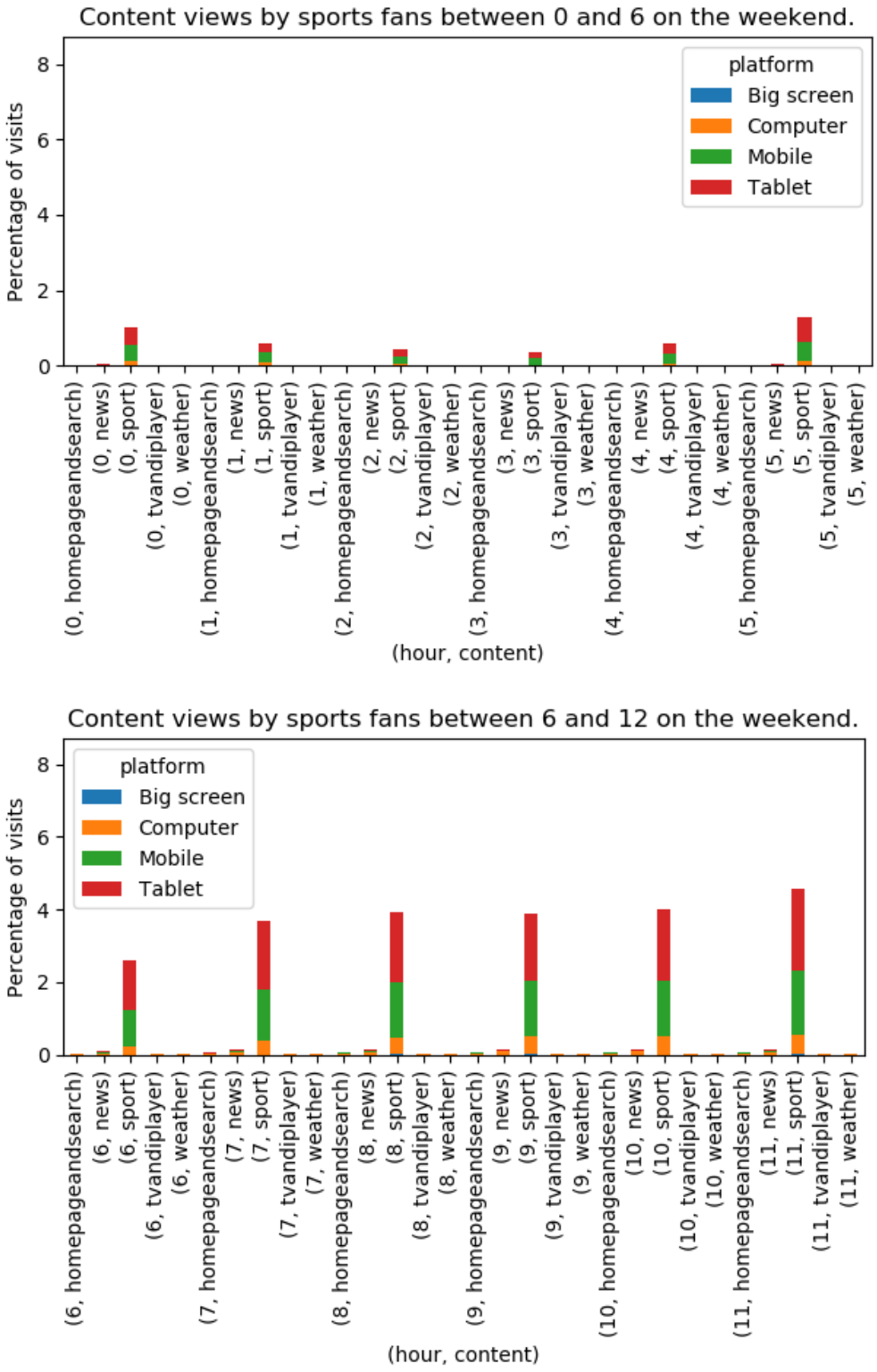

Figure 4.3: Sports fans' content views via platform from midnight to 6am (above) and from 6am to midday (below) on the weekend.

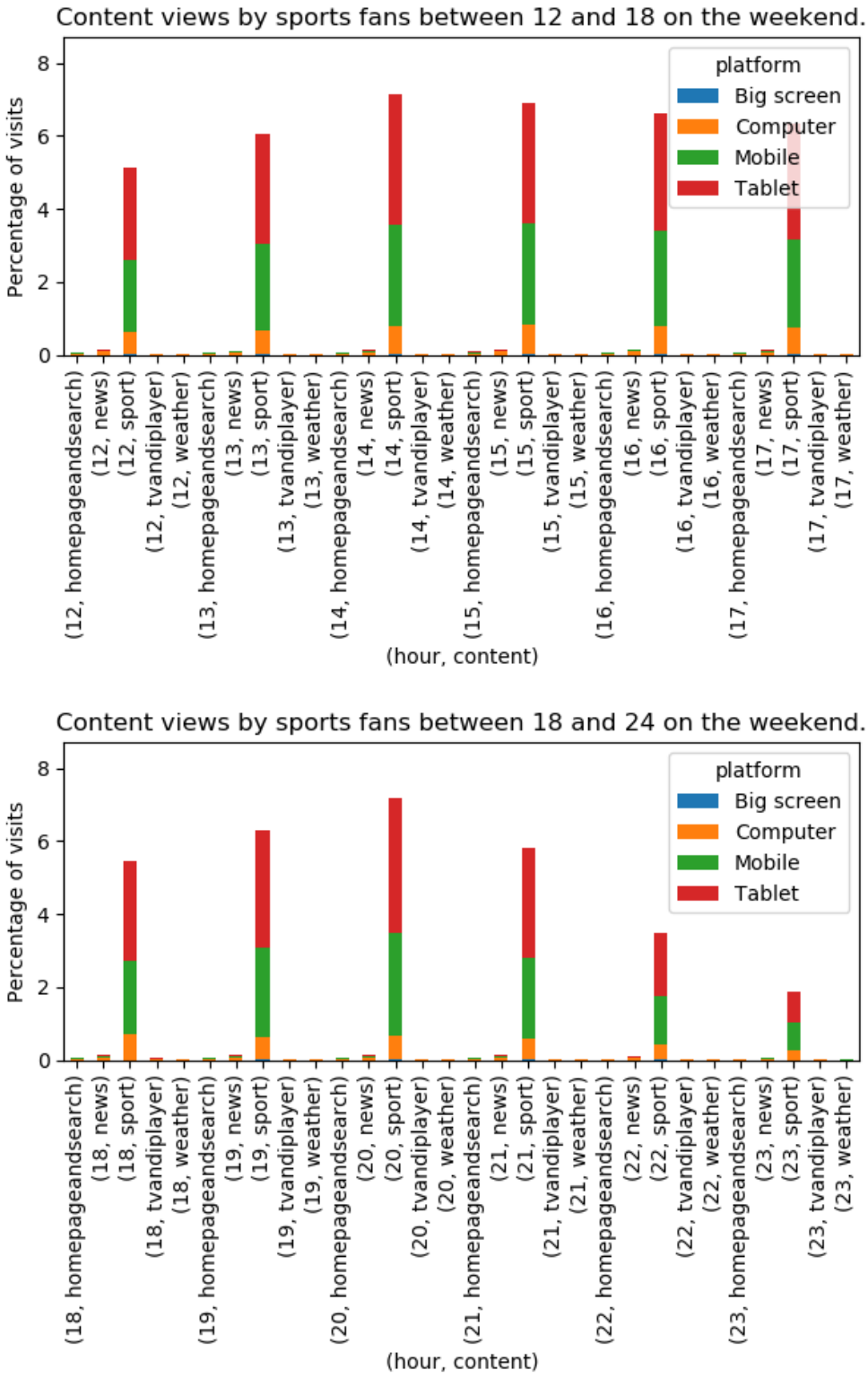

Figure 4.4: Sports fans' content views via platform from midday to 6pm (above) and from 6pm to midnight (below) on the weekend.

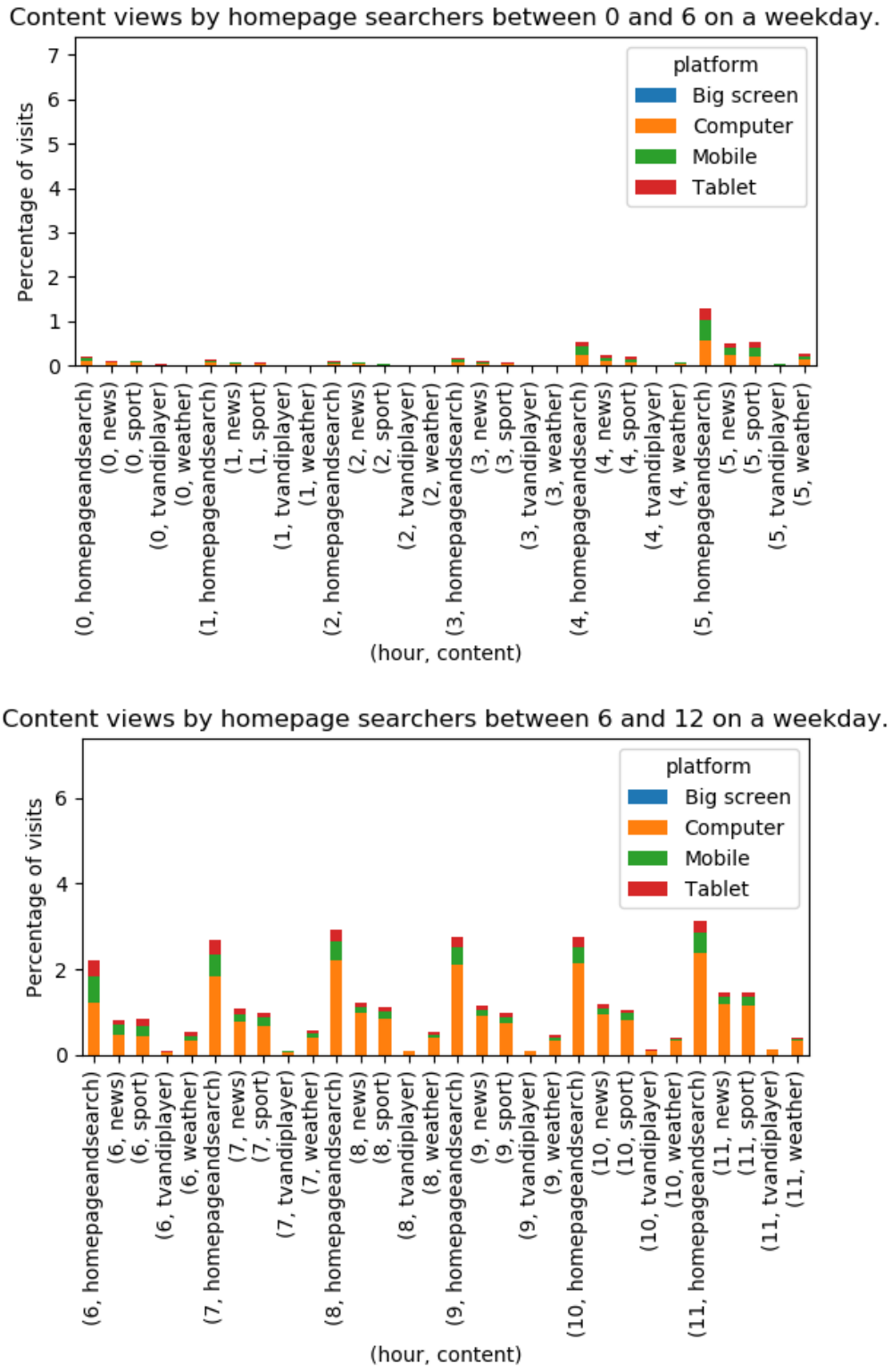

Figure 4.5: Homepage searchers' content views via platform from midnight to 6am (above) and from 6am to midday (below) on a weekday.

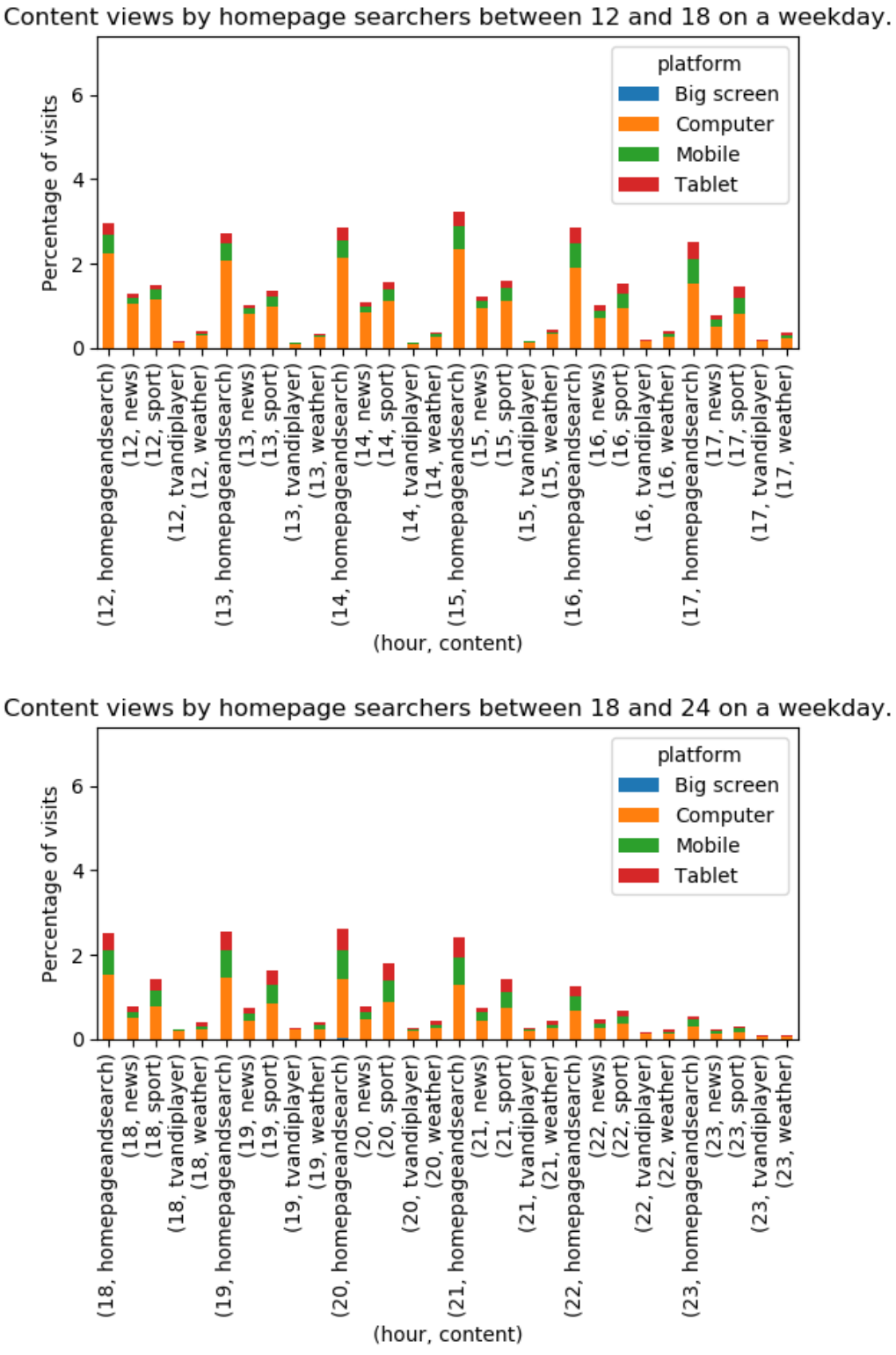

Figure 4.6: Homepage searchers' content views via platform from midday to 6pm (above) and from 6pm to midnight (below) on a weekday.

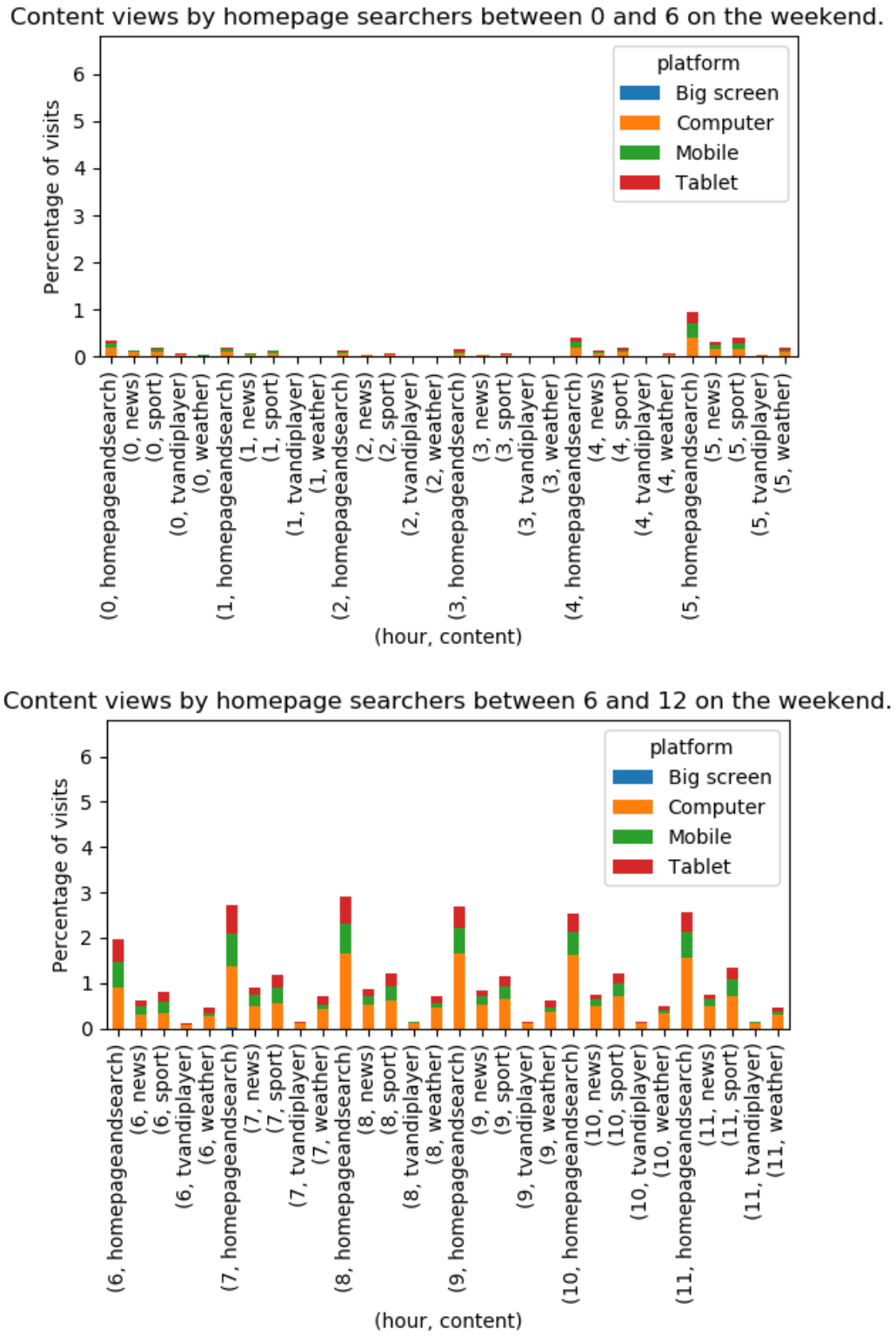

Figure 4.7: Homepage searchers' content views via platform from midnight to 6am (above) and from 6am to midday (below) on the weekend.

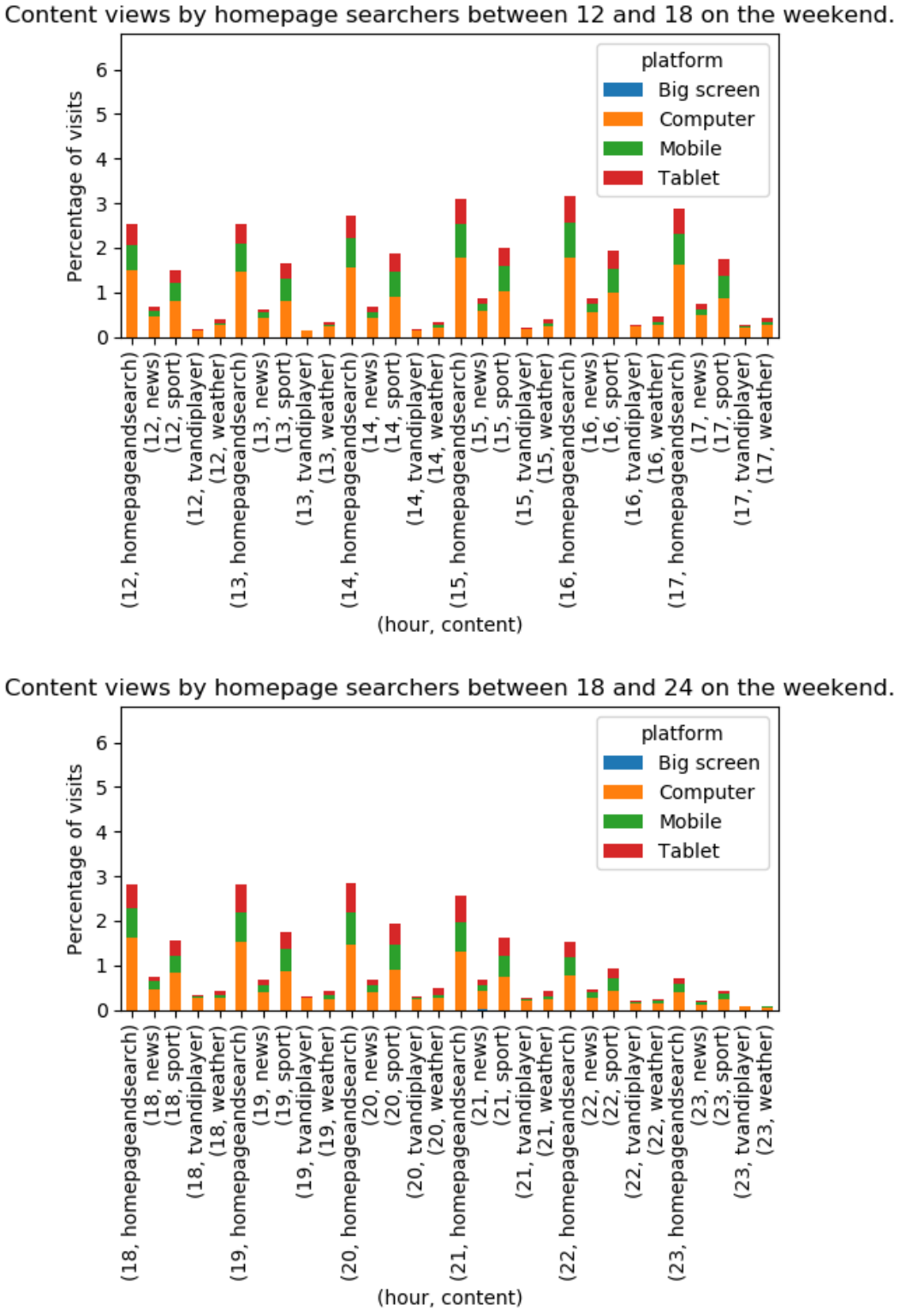

Figure 4.8: Homepage searchers' content views via platform from midday to 6pm (above) and from 6pm to midnight (below) on the weekend.

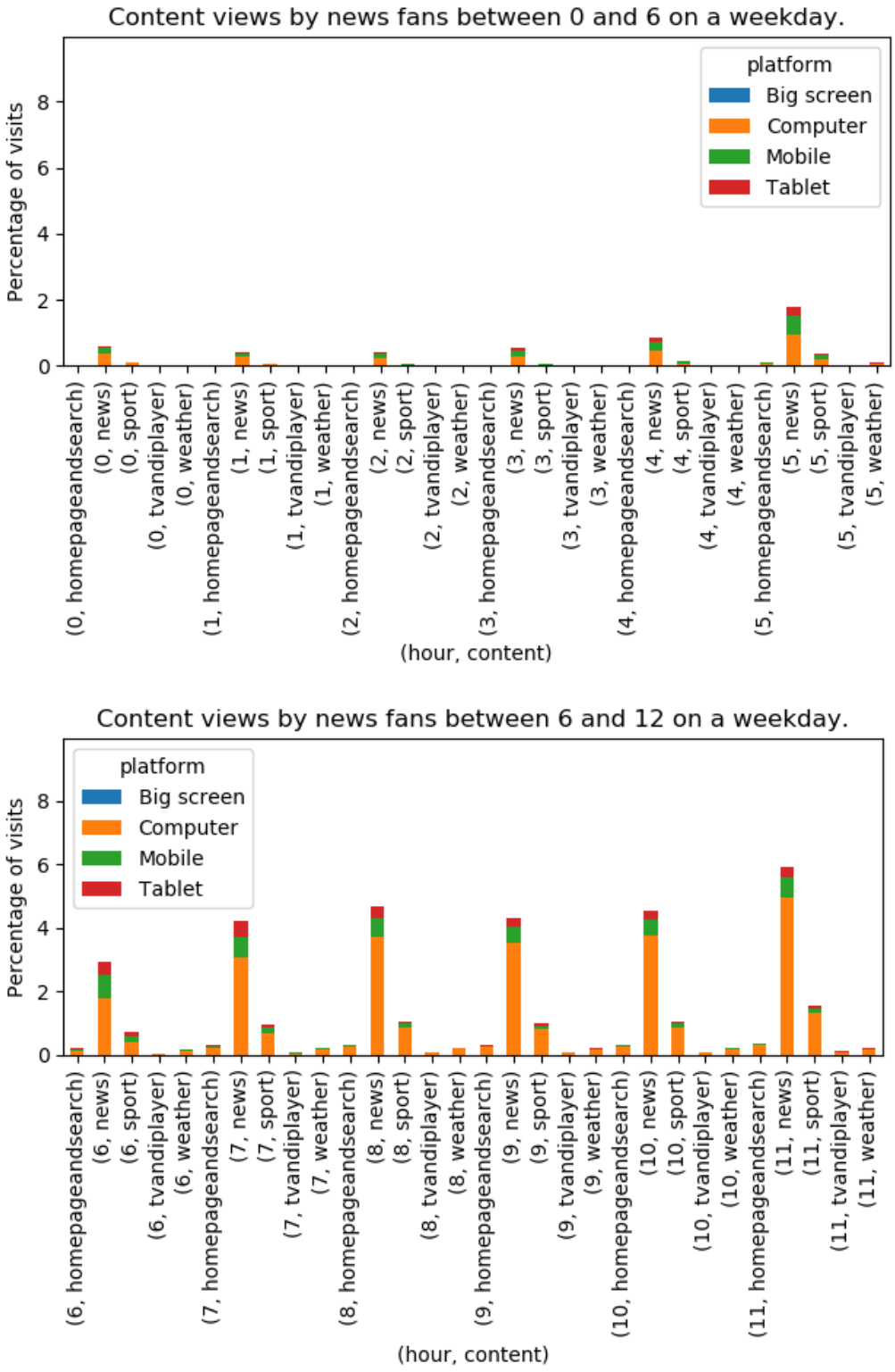

Figure 4.9: News fans' content views via platform from midnight to 6am (above) and from 6am to midday (below) on a weekday.

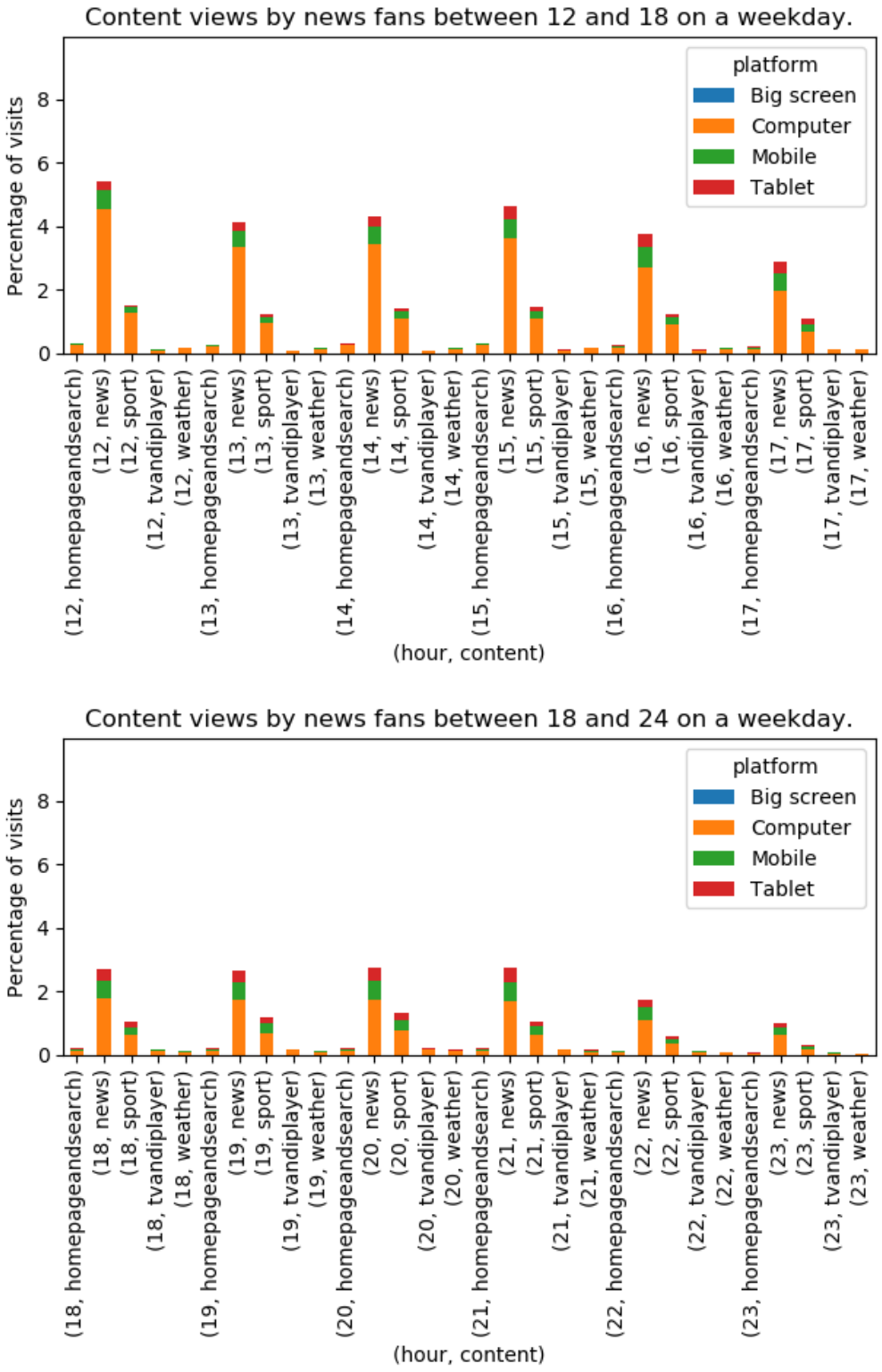

Figure 4.10: News fans' content views via platform from midday to 6pm (above) and from 6pm to midnight (below) on a weekday.

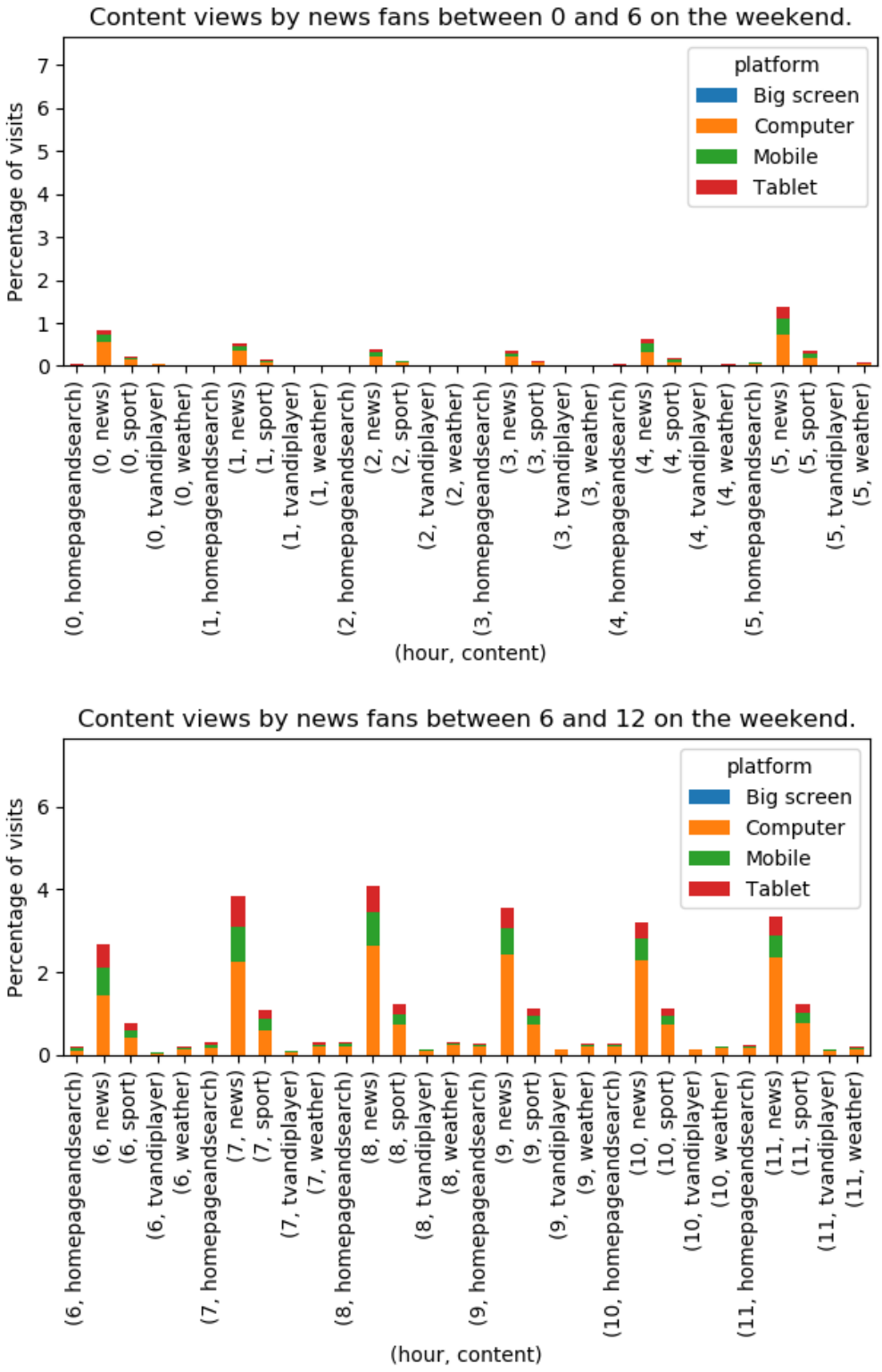

Figure 4.11: News fans' content views via platform from midnight to 6am (above) and from 6am to midday (below) on the weekend.

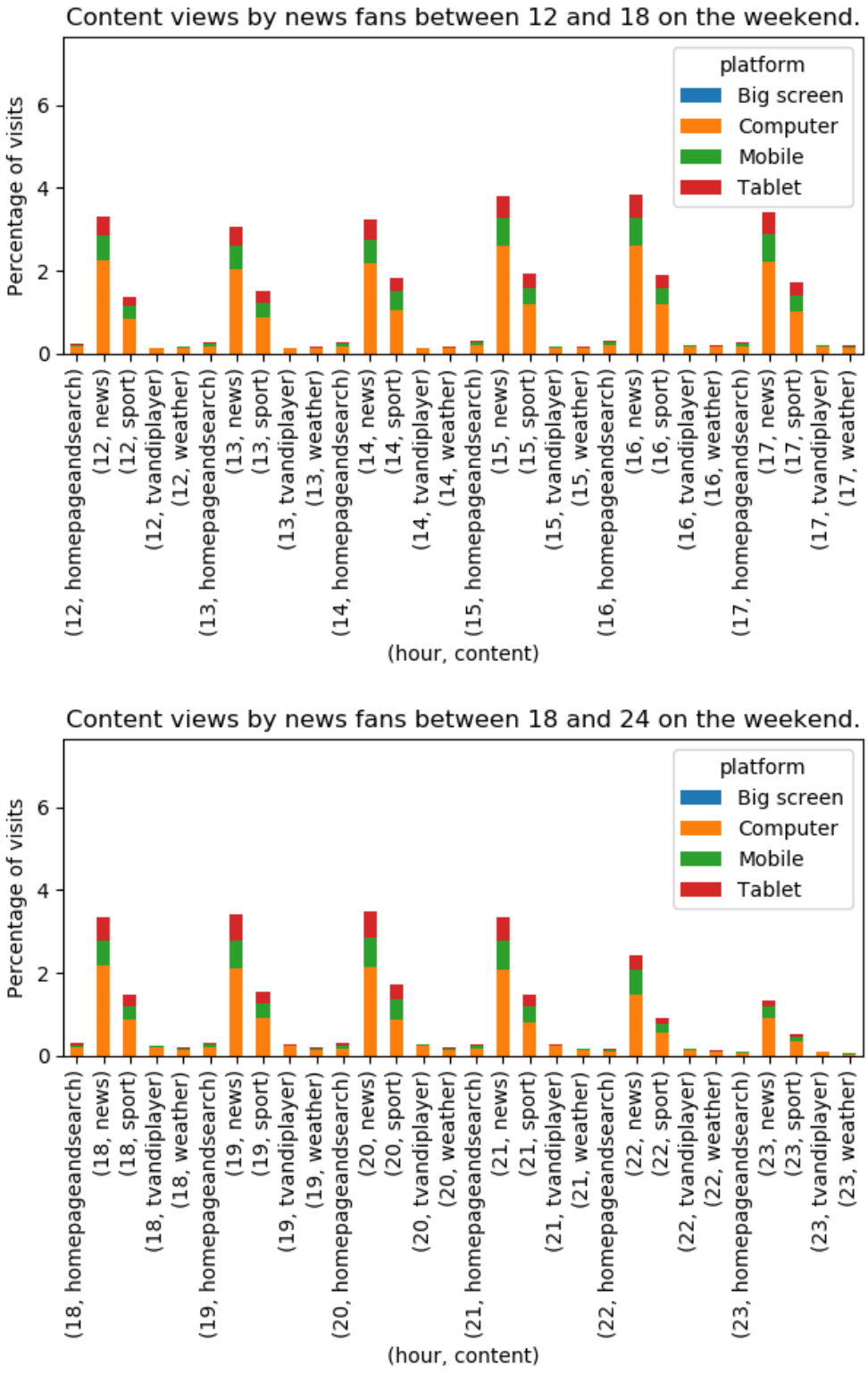

Figure 4.12: News fans' content views via platform from midday to 6pm (above) and from 6pm to midnight (below) on the weekend.

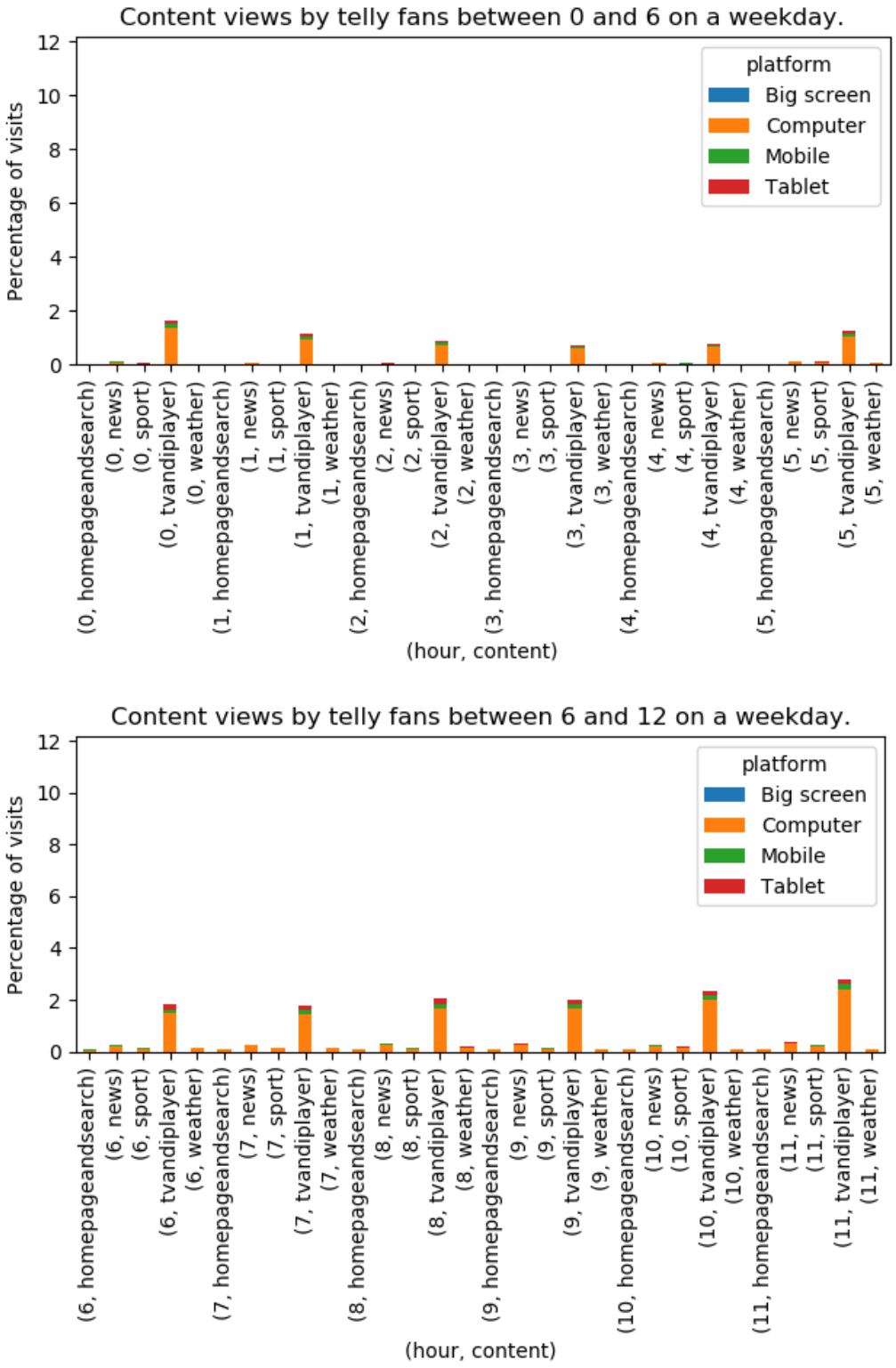

Figure 4.13: Telly fans' content views via platform from midnight to 6am (above) and from 6am to midday (below) on a weekday.

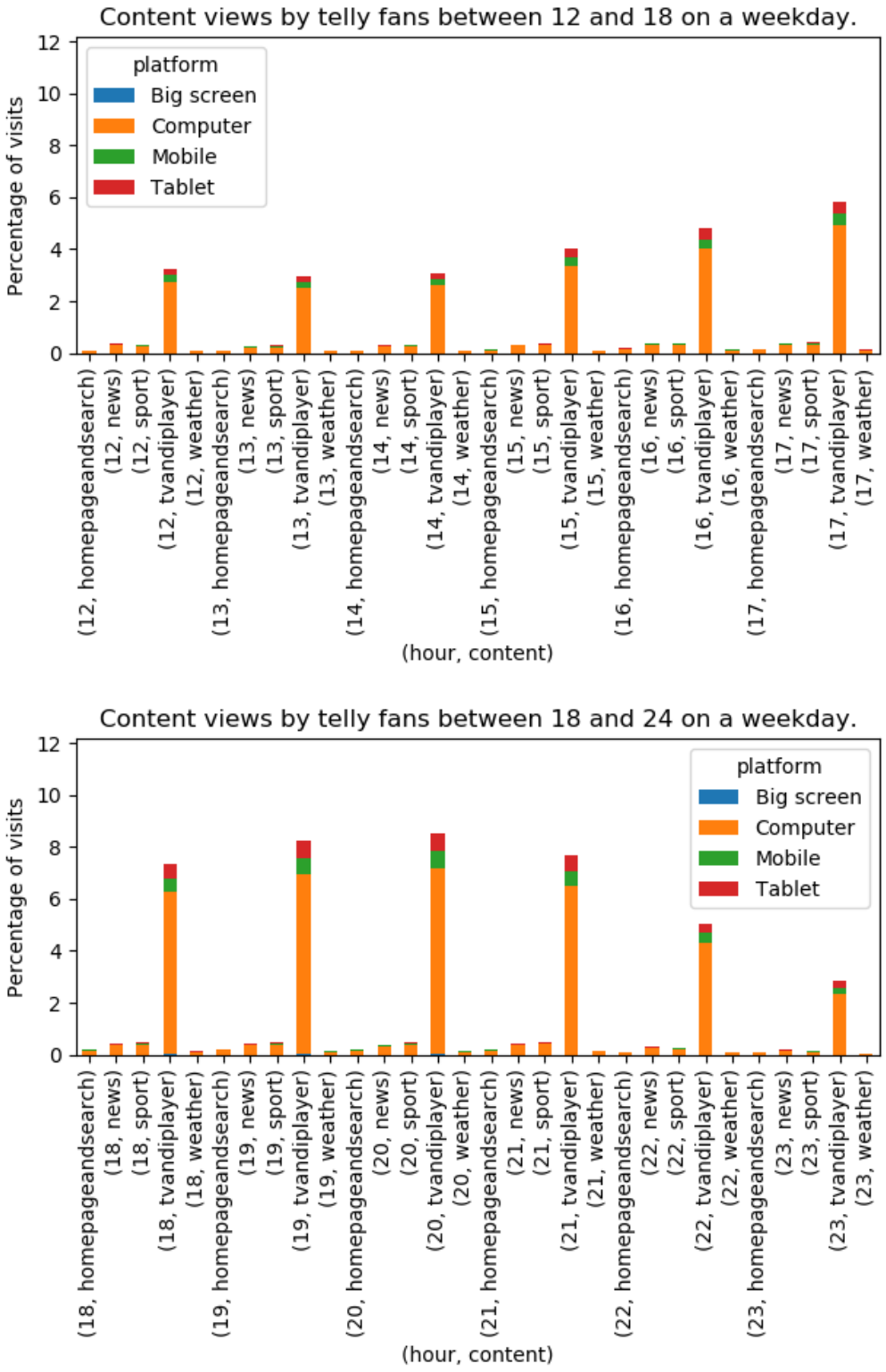

Figure 4.14: Telly fans' content views via platform from midday to 6pm (above) and from 6pm to midnight (below) on a weekday.

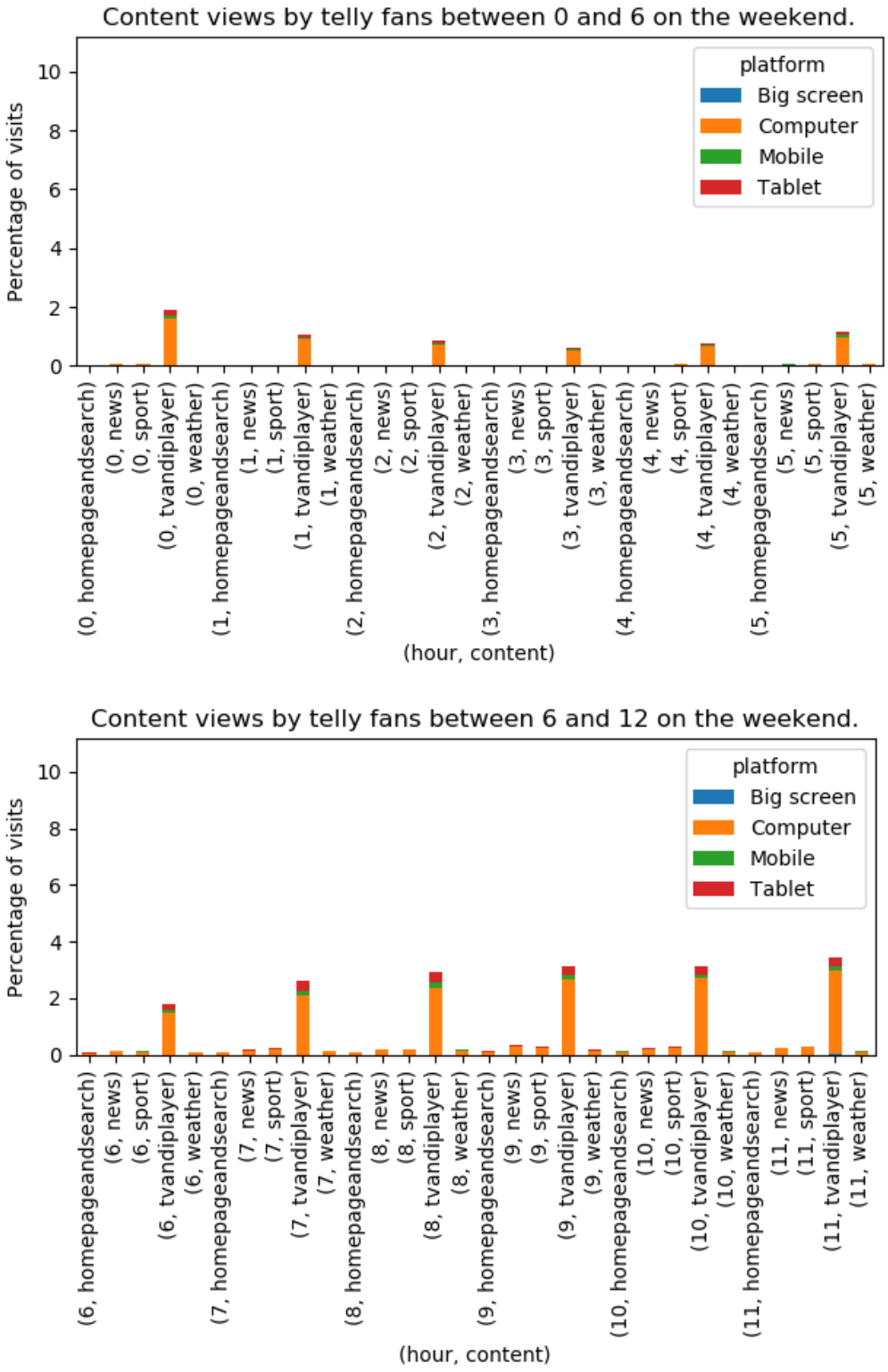

Figure 4.15: Telly fans' content views via platform from midnight to 6am (above) and from 6am to midday (below) on the weekend.

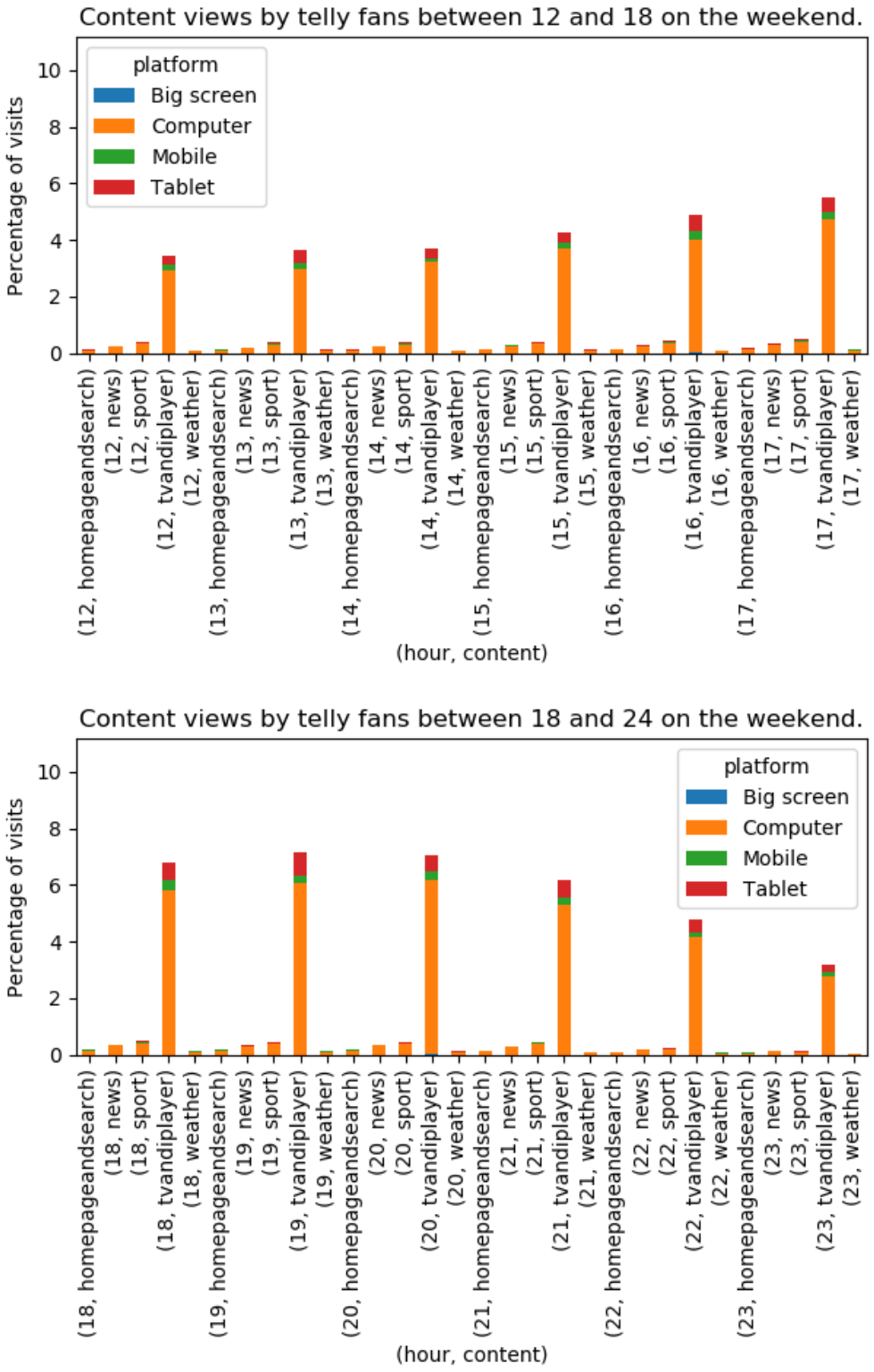

Figure 4.16: Telly fans' content views via platform from midday to 6pm (above) and from 6pm to midnight (below) on the weekend.