User Engagement in Online News: Under the Scope of Sentiment, Interest, Affect, and Gaze

Ioannis Arapakis, Mounia Lalmas, B. Barla Cambazoglu
Mari-Carmen Marcos, Joemon M. Jose

August 22, 2013

Abstract

Online content providers, like news portals and social media platforms, constantly seek new ways to attract large shares of online attention by keeping their users engaged. A common challenge is to identify which aspects of online interaction influence user engagement the most. In this article, through an analysis of a news article collection obtained from Yahoo! News, we demonstrate that news articles exhibit considerable variation in terms of the sentimentality and polarity of their content, depending on factors like news provider and genre. Moreover, through a laboratory study, we observe the effect of sentimentality and polarity of news and comments on a set of subjective and objective measures of engagement. In particular, we show that attention, affect, and gaze differ across news of varying interestingness. As part of our study, we also explore methods that exploit the sentiments expressed in user comments to reorder the lists of comments displayed in news pages. Our results indicate that user engagement can be anticipated if we account for the sentimentality and polarity of the content, as well as other factors that drive attention and inspire human curiosity.

1 Introduction and Background

The proliferation of digital media, coupled with the phenomenal growth of online content, has spur the demand for diverse ways of searching, locating and accessing information. A similar trend has been also witnessed in news consumption, which is progressively transcending the boundaries of printed press into the area of online news content distribution (news portals, blogs, micro-blogs, social networks), and is becoming increasingly interactive and social. According to the Newspaper Association of America, 74% of all Internet users rely on local newspaper media, digital as well as print, as sources of information [22]. Considering that the average advertising revenue of a newspaper ranges between 60% to 70% of the total revenue [27], depending on the circulation, news portals constantly seek new ways to keep their users engaged by serving them interesting content in an attractive and enticing manner.

In general, interest in online content is highly asymmetrical; only a small portion of that content receives significant attention, while the remaining majority is barely noticed [39]. But, what really drives user interest and facilitates engagement with the news content? This is the question we attempt to answer in this article. In particular, we examine which aspects of user interaction influence engagement the most in the online news domain, while considering the fundamental role of user interest. In addition, we demonstrate how standard engagement metrics and content-based sentiment features can be leveraged to quantify the level of interaction between the user and the news content. Previous studies have mostly examined the relationship between information propagation and user supplied comments [45, 40, 46] or web content properties [4, 6, 8, 50, 53, 55].

One of the characteristics of user engagement is the affective dimension of the interaction between the user and the content. A potential way of examining this dimension is analysing the sentiments expressed within the content. So far, different aspects of sentiment analysis have been investigated by the research community, such as extraction [1, 3, 54], classification [13, 34, 47], retrieval [17, 56], summarization [7, 25], and presentation [19]. Sentiment analysis techniques mainly focused on identifying the linguistic features that contribute to the affective content of text and how these features could be automatically extracted to obtain a sentiment metric. In particular, in [24], two complementary sentiment metrics were defined: sentimentality, which quantifies the total amount of sentiments expressed in the content, and polarity (attitude), which represents the inclination towards positive or negative sentiments, i.e., the direction of aggregated sentiments (Section 2.2 contains more details about our use of these metrics). While sentiment analysis has been applied in news and blog analysis [5, 14, 18, 33, 50, 55], the interaction between sentiment and user engagement in online news content remains largely unexplored.

In this work, one of our goals is to bridge the divide between sentiment metrics that derive from the content and offline behaviour measures of engagement. We aim to establish which aspects of user interaction influence engagement the most in the online news domain. To this end, we conducted an analysis of 13,319 news articles taken from Yahoo! News US.[1] Our analysis demonstrates that news articles exhibit considerable variation in terms of the sentimentality and polarity of their content, depending on variables like news provider and genre. Given that variation, we employed a smaller subset of our news collection and observed the real-life effects of articles' sentiment characteristics on the news reading experience, through a correlation analysis.

As another contribution, we employed a set of user engagement measures [2], namely positive and negative affect, focused attention, and gaze behaviour, and demonstrated how they differ across online news of varying interestingness. In general, engagement measures can be divided into two broad types: subjective and objective. Subjective measures are central to user engagement and generally involve self-reports or interviews that rely heavily on user's perception. Thus, they are susceptible to post-hoc interpretation bias. On the other hand, objec-

[1] `http://news.yahoo.com/us/`.

tive measures like neuro-physiological signals and web analytics can provide a more reliable and objective mapping of user actions, independent of user's perceptual ability, but it is not always clear which specific aspects of engagement they target. Our approach to studying engagement is holistic in the sense that it leverages the limitations with each type of measure. In particular, we observed that the experienced affect and the level of attention are determined to a large extent by user interest, which plays a fundamental role in user engagement. We also discovered that, although interestingness is user-dependent, it can be anticipated to some degree if we account for factors like the sentimentality and polarity of the content.

Finally, we exploited the sentiments expressed in user comments to reorder the list of comments displayed in news articles and observed the effect of different comment orderings on user engagement. So far, comments have been examined for the most part in terms of volume [40, 46] or ratings (e.g., likes, votes), which would be one way of measuring engagement. For example, the number of comments posted in an article is considered as a measure of popularity or user participation. Our approach is broader and treats comments as content rich in sentiment information, which can provide useful insights about people's experiences on a particular activity domain [49]. We show that certain comment orderings are more preferable than others depending on article interestingness and user gender.

The rest of the paper is organised as follows. In Section 2 we present our analysis of online news articles and demonstrate how they vary in terms of the sentimentality and polarity of their content. In Section 3 we provide details about our controlled experimental user-study, where we jointly analyse the affective and attention components of engagement with the content-based features of news articles, on a subset of our news collection. In Section 4 we review our results and in Section 5 discuss our main findings. Finally, we conclude the paper and suggest future research directions in Section 6.

2 Sentiment Analysis of News Articles

We begin our analysis with a study of news articles taken from Yahoo! News. The goal of this analysis is to demonstrate that news articles exhibit significant variation in terms of the sentimentality and polarity of their content, depending on variables like news provider and genre.

2.1 News Dataset

Our dataset contained 13,319 news articles from Yahoo! News US, all written in English. Yahoo! News is a news portal and a news aggregator by Yahoo! that categorises news into "Top Stories", "U.S. National", "World", "Business", "Entertainment", "Science", among several other genres. Articles in Yahoo! News come from news services such as the Associated Press, CNN, BBC, and Reuters.

The Yahoo! News portal is available in different regions and languages, including Brazil, France, Germany, India, Spain, United Kingdom, and US.

Our dataset was constructed by crawling news articles from the US portal of Yahoo! News, over a period of two weeks. We connected to the RSS feed of Yahoo! News every 15 minute and discovered newly published articles. We then fetched these articles from Yahoo! News. Each article was identified by its unique URI and stored in a database, along with some meta-data, such as article's genre (e.g., politics, sports, crime), its publication date, and its HTML content on the day of the publication.

2.2 Sentiment Analysis Tool

To compute the degree (sentimentality) and sign (polarity) of the sentiments expressed in the news articles and user comments, we used SentiStrength, a widely used lexicon-based, sentiment analysis tool [42]. SentiStrength was designed especially for sentiment detection in short informal text and has been deployed and evaluated successfully in several studies [24, 41, 42, 43, 44, 48, 52]. For a given piece of short text, the SentiStrength tool generates a positive and a negative sentiment score for each word in the text. When computing these two word-level scores, SentiStrength relies on a number of editorially created word lists, including a sentimental word list, a booster word list, an idiom list, a negation word list, and an emoticon list. The generated positive scores are integers in the range from $+1$ (neutral) to $+5$ (extremely positive), while the negative scores are integers in the range from -1 (neutral) to -5 (extremely negative). The tool returns the maximum/minimum word scores of the positive/negative sentiment scores associated with the text. For example, the sentence "*President Obama tries to reform food aid as agribusiness fights back*" would receive the following classification rationale: President Obama tries to reform food aid[+2] as agribusiness fights[−3] back. In the above sentence, the word "aid" has received the maximum (within the sentence) positive sentiment score $+2$, whereas the word "fights" has received the maximum negative sentiment score -3. These two scores characterise the sentimentality/polarity of that particular sentence.

We used the positive/negative scores returned by SentiStrength to compute the sentimentality and the polarity of a sentence. Following [24], we computed the sentimentality of a sentence as the sum of the absolute values of the positive and negative scores associated with the sentence. We determined the polarity of the sentence as a sum of its positive and negative scores. The sentimentality of a news article was then computed by taking an average over the sentimentality scores of the sentences in the article. The polarity of the article was computed in a similar manner by averaging the polarity scores of the sentences in the article.

2.3 Data Characterisation

Herein, we analyze the change in sentimentality and polarity of news content depending on i) the news source the article is obtained from, ii) the genre of the article, and iii) the time the article is published on the Web.

News provider: As discussed in Section 2.1, the news articles in Yahoo! News come from various news providers, like the Associated Press and Reuters. In Figure 1, these news providers are represented by different circles, and the size of each circle indicates the volume of news articles generated by the corresponding provider. The sentimentality and polarity scores of a provider were computed by averaging the sentimentality and polarity scores of all articles obtained from the provider, respectively. The articles in our collection were mainly obtained from two big news providers, Reuters and Associated Press. The remaining providers were relatively small in terms of their contribution to the published articles. According to Figure 1, news sources show variation in terms of the average sentimentality and polarity of the articles they provide. The two big news providers are somewhat similar to each other in that both provide relatively more sentimental and negative news articles, while Reuters provides slightly more sentimental articles than Associated Press. We observe several small-scale news providers that provide positive or neutral news articles. These are mainly some "niche" news providers that serve technical news articles or press releases.

Genre: The genre of an article is one of the main factors that affect the sentimentality and polarity of the article. Our news collection has a fine-grain categorization of news articles into genres, performed by professional editors. In our work, we further combined some related genres and came up with a more coarse-grain categorization, including 14 genres. Figure 2 shows the average sentimentality and polarity of the articles belonging to these genres. According to this figure, there is high variation depending on the genre. For articles in genres like law, health, entertainment, politics, and life, we observe a larger amount of sentimentality. On the other hand, genres like technology and science have the lowest sentimentality. In terms of polarity, the news articles are already known to be negative compared to other types of media content [53]. In our case, we observe the most negative articles in genres such as law, health, and politics. The entertainment news are much less negative although they have high sentimentality.

We also examined the differences in the sentimentality and polarity of news titles and articles across the available genres. The Kruskal-Wallis test was applied to analyse the observed intra-genre variability. In all cases, the title sentimentality ($H(13) = 693.63, p < .001$), the title polarity ($H(13) = 932.13, p < .001$), the body of article sentimentality ($H(13) = 1330.40, p < .001$), and the body of article polarity ($H(13) = 3394.69, p < .001$) were significantly affected by the article genre. Mann-Whitney tests were used to follow up these findings for all pair-wise comparisons of news genres. The effect sizes are presented in Table 1. In this analysis, a Bonferroni correction was applied and so all effects were examined at a 0.005 level of significance. Those that were not found statistically significant were omitted. We also followed Cohen's [10, 11] suggestion on what constitutes a large or small effect. The effect sizes shown in Table 1 validate our initial assumption that news genres vary significantly on the sentimentality

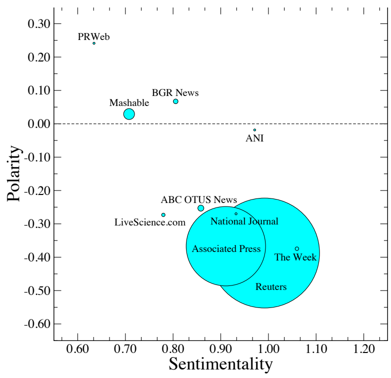

Figure 1: Variation of news volume, sentimentality, and polarity with respect to the news provider.

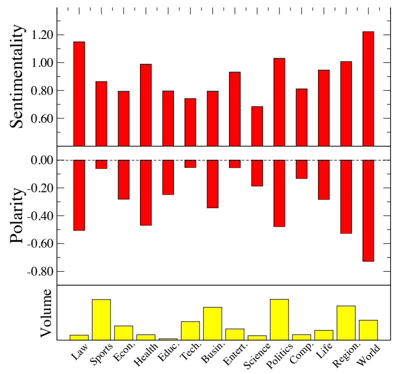

Figure 2: Variation of news volume, sentimentality, and polarity with respect to the genre.

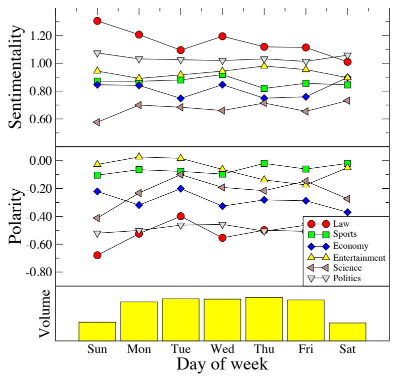

Figure 3: Variation of news volume, sentimentality, and polarity with respect to the day of the week.

and polarity characteristics of their content, potentially creating diverse news reading experiences.

Sentiment-rich content is more attractive and can provide useful insights about the general sentiment tendency towards a topic. This theory is partly supported by our findings, which suggest that web users are drawn to highly sentimental news that are characterised by a negative polarity. This makes sense since news headlines are written to induce emotional reactions and attract people's attention, and news articles are already known to be negative compared to other types of media content [53].

Time: We next investigate whether the time at which an article is published has an effect on the article's sentimentality or polarity. We grouped articles according to the hour of the day and the day of the week in which they are published, to see if there is any variation in our metrics. In general, we did not observe a significant variation based on time. In Figure 3, we report the results only for the day of the week and a selected set of genres. We do not pursue this further in this paper; we nonetheless acknowledge this results, which may be of interest to others.

In this section we presented our analysis of a news articles collection, taken from Yahoo! News US, and demonstrated that news exhibit a considerable variation in terms of the sentimentality and polarity of their content. The degree (sentimentality) and sign (polarity) of the sentiments expressed in the news articles were computed using SentiStrentgh, a widely used sentiment analysis tool. Our initial findings indicate that the sentimentality and polarity depend on factors such as the news source and news genre, whereas the news content is portrayed as highly-sentimental and of negative polarity. In the following section we study the real-life effects of articles' sentiment characteristics on the news reading experience, which a focus on user engagement. We also employ a set of subjective and objective measures of engagement and demonstrate how they differ across online news of varying interestingness.

3 User Study Design

Given the variation observed in the sentiment features of online news, we employed a subset of our news collection and investigated the effects of sentimentality, polarity, and news content of varying interestingness, on user engagement. Our lab-based approach involved a controlled experimental setting that allowed us to jointly analyse the affective and attention characteristics of engagement with the content-based features (sentimentality and polarity) of news articles. We examined positive and negative affect, focused attention, and gaze behaviour. The behaviour data were captured using an eye tracker, while qualitative information was collected using questionnaires. We describe the subjective (affect, focused attention) and objective (gaze behaviour) measures of engagement we employed.

3.1 Dataset

We limited our user study to articles selected from three genres: (i) crime and law, (ii) entertainment and lifestyle, and (iii) science. The frequency distribution of the word count of the selected articles followed a bimodal pattern, with the bulk of the articles located around the mean score 447.47. Using that value as a reference point, we applied a $+/-150$ word limit and filtered out the articles that contained less than 300 or more than 600 words; thus ensuring that all articles in our collection were approximately of the same length. From the resulting 383 articles, we took a random sample of 40 articles per genre. Twenty-four volunteers rated this sample on a 5-point interestingness scale. The obtained

Table 1: Effect sizes from the analysis of variance of news genres with respect to the sentimentality and polarity of news articles

	Business	Computers	Economy	Education	Entertainment	Health	Law	Life	Politics	Regional	Science	Sports	Tech	World
Sentimentality														
Business		-0.09**	-0.10**	-0.10**	-0.13**	-0.26**	-0.20**	-0.31**	-0.30**	-0.17**	-0.23**	-0.04**	-0.04*	-0.37*
Computers			-	-0.12*	-	-0.27**	-0.17**	-0.27**	-0.13**	-	-0.26**	-	-0.08**	-0.24**
Economy				-0.08*	-0.05*	-0.24**	-0.16**	-0.28**	-0.18**	-0.05**	-0.22**	-0.04*	-0.08**	-0.30**
Education					-	-0.11*	-	-0.10*	-	-	-	-0.06**	-0.12**	-0.08**
Entertainment						-0.18**	-0.11**	-0.21**	-0.11**	-	-0.16**	-0.08**	-0.11**	-0.22**
Health							-0.09*	-	-	-0.14**	-	-0.18**	-0.29**	-
Law								-0.11**	-	-0.08**	-	-0.13**	-0.22**	-0.10**
Life									-0.05**	-0.18**	-0.07*	-0.23**	-0.34**	-
Politics										-0.14**	-	-0.24**	-0.26**	-0.08**
Regional											-0.11**	-0.10**	-0.13**	-0.22**
Science												-0.16**	-0.28**	-0.07*
Sports													-	-0.30**
Tech														-0.37**
World														
Polarity														
Business		-0.15**	-	-	-0.26**	-0.19**	-0.12**	-0.03*	-0.25**	-0.17**	-	-0.29**	-0.31**	-0.43**
Computers			-0.18**	-0.16**	-0.07*	-0.41**	-0.34**	-0.20**	-0.26**	-0.24**	-0.17**	-	-0.06*	-0.46**
Economy				-	-0.27**	-0.23**	-0.15**	-	-0.21**	-0.15**	-	-0.24**	-0.29**	-0.43**
Education					-0.17**	-0.20**	-0.13*	-	-0.08**	-0.06**	-	-0.09**	-0.15**	-0.21**
Entertainment						-0.43**	-0.37**	-0.28**	-0.37**	-0.35**	-0.21**	-0.03*	-	-0.57**
Health							-0.08*	-0.17**	-	-0.06**	-0.24**	-0.31**	-0.41**	-0.14**
Law								-0.10**	-0.04**	-	-0.16**	-0.25**	-0.35**	-0.21**
Life									-0.13**	-0.08**	-	-0.21**	-0.27**	-0.32**
Politics										-0.07**	-0.13**	-0.47**	-0.45**	-0.19**
Regional											-0.10**	-0.41**	-0.42**	-0.11**
Science												-0.14**	-0.20**	-0.31**
Sports													-0.03**	-0.59**
Tech														-0.63**
World														

*. Statistical significance at the .05 level (2-tailed).
**. Statistical significance at the .001 level (2-tailed).

scores allowed us to determine the three most interesting and the three least interesting news articles for each genre (18 articles in total).

For each news article, we crawled the first 100 comments that appeared in the available comment orderings “popular now”, “newest”, and “most replied”. The comments in “popular now” ordering were sorted in decreasing order of a popularity factor, which was determined by the likes/dislikes that each comment received. The comments in the “newest” category were sorted in increasing order according to their posting date, while the comments in “most replied” were sorted in decreasing order of the number of replies they had received. We also introduced four new orderings: (i) “sentimentality high”, (ii) “sentimentality low”, (iii) “polarity (+)”, and (iv) “polarity (−)”. The “sentimentality high” and “sentimentality low” orderings listed the comments in decreasing or increasing order of the sentiment scores computed with SentiStrength, respectively.[2] Similarly, “polarity (+)” and “polarity (−)” orderings sorted the comments according to their polarity. These additional orderings allowed us to observe the effect of different approaches to exploiting the sentiments expressed in user comments on the news reading experience and, consequently, on user engagement.

3.2 Measures of Engagement

We examined engagement both as a process and as an outcome of an online news reading experience. Given its multi-faceted nature, we employd a set of subjective and objective measures, such as positive and negative affect, focused attention, and gaze behaviour:

Affect refers to the emotion mechanisms that influence our everyday interactions and can act as the primary motivation for sustaining our engagement [28, 30, 51] during an information processing task or a computer-mediated activity. Focused attention [32] is also central to engagement; a feeling of energised focus and total involvement, often accompanied by loss of awareness of the outside world and distortions in the subjective perception of time. Gaze behaviour is also known to be associated with attention, interest, and perception, and has been extensively studied in information processing tasks like reading [9], visual search [35], scene perception [21], and recently in micro-blogging [12]. The importance of gaze in the assessment of engagement lies in the fact that, although looking might appear to be a process that is under voluntary control, conscious and deliberate control of fixation happens infrequently. As with other components of voluntary performance (e.g., walking or maintenance of balance), looking is controlled by a general intention, and consciousness plays a minor role in the execution of the intended sequence of fixations [23].

3.3 Questionnaires

We used two types of questionnaires. The first questionnaire (entry) was used at the beginning of the study to determine which news articles were perceived

[2] The sentimentality and polarity of a user comment is computed by averaging the individual values associated with the sentences in the user comment, as discussed in Section 2.2.

Table 2: PANAS [51]

Positive Affect items		Negative Affect items	
active	excited	afraid	irritable
alert	inspired	ashamed	jittery
attentive	interested	distressed	nervous
determined	proud	guilty	scared
enthusiastic	strong	hostile	upset

as most interesting or most uninteresting by the participants, prior to reading the articles. The second questionnaire (main) was administered during the task and gathered information on the experienced affect, focused attention, degree of interest after reading the assigned articles, as well as other information like demographics. The questions appeared in a random sequence to prevent a potential bias due to the ordering effect, and inquired about the following aspects:

PANAS: The Positive and Negative Affect Scale [51] was used to measure the affect before and after each task (Table 2). PANAS is a standardised test for measuring long-term affect changes. It includes 10 items measuring positive affect (PAS) and 10 items measuring negative affect (NAS). Participants were asked to respond on a 5-point Likert scale (very slightly or not at all; a little; moderately; quite a bit; extremely) their agreement to the statement: "You feel this way right now, that is, at the present moment", for each item. Affect was also measured by asking the participants to respond to the question "Overall, did you feel positive or negative while completing the news reading task?".

Although PANAS may not be as efficient and accurate for capturing temporal micro-resolutions of emotional responses, there are several examples of studies from the domain of Library & Information Science [20, 26, 30, 32] where PANAS has been successfully applied for measuring searchers' affect between search tasks. Considering that the duration of our news reading tasks is comparable to those in the aforementioned studies, we believe that our experimental approach to measuring emotion was reasonably accurate.

Focused Attention: A 9-item focused attention (FA) subscale, part of a larger scale for measuring user engagement [32], was adapted to the context of the news reading tasks. The FA scale has been used in past work [28, 30, 31] to evaluate users' perceptions of time passing and their degree of awareness about what was taking place outside of their interaction with the task at hand. For our news reading task, participants were instructed to state on a 5-point Likert scale (strongly agree; disagree; neither agree nor disagree; agree; strongly agree) their agreement to each item shown in Table 3. Additionally, the perceived time spent on the news reading task was measured using the question "Please estimate in minutes the total time you think you spent reading the news article and the comments, without including the time spent reading and answering the survey questions", and was compared against the actual time spent on the task.

Table 3: Focused attention scale [32]

1. I forgot about my immediate surroundings while performing this news task.
2. I was so involved in my news task that I ignored everything around me.
3. I lost myself in this news task experience.
4. I was so involved in my news task that I lost track of time.
5. I blocked out things around me when I was completing the news task.
6. When I was performing this news task, I lost track of this world around me.
7. The time I spent performing the news task just slipped away.
8. I was absorbed in my news task.
9. During this news task experience I let myself go.

Interest: Given that the participants were asked to rank the news articles based on the interestingness of their titles, prior to the completing the task, we wanted to validate the effectiveness of our experimental manipulation. Therefore, participants' level of interest in the news articles was also measured post-task on a 5-point Likert scale, using the following set of questions: "I found the news article interesting to read", "I was familiar with the content of the news article", "The news article was easy to understand", "I enjoyed reading the news article", "I typically read this type of news on the Web", "I wanted to find out more about the news I read".

Comments: The different orderings of associated comments were evaluated on a 5-point Likert scale by the following five questions: "The comments I read were interesting", "I found the comments useful in understanding better the article", "The comments provided additional insight for the article", "The comments improved my news reading experience", "I enjoyed reading these comments".

Demographics: This section gathered demographics and background information on the news reading habits of the participants. A set of questions was developed specifically for our study, such as "How often do you read the news online?", "On average, how much time in minutes do you spend reading the news online, in a given day?", "How often do you leave comments in the news articles you read online?", and "How often do you browse other people's comments in the news articles you read online?", which allowed us to determine participants' previous experience with online news and social media.

3.4 Eye Tracking

The relationship between attention and eye movements has been investigated extensively in the past [15, 36, 37, 38]. Gaze direction has been considered an indicator of the focus of attention. When we read, examine a scene, or search for an object, we continuously make eye movements called saccades. Saccades are rapid movements that occur when we change focus, and can reach velocities as high as 500° per second. When the visual gaze is maintained on a single location

Table 4: Eye metrics used to analyse gaze behaviour

1. Time to First Fixation: Time taken (in seconds) before a participant fixates on an AOI for the first time.
2. Fixations Before: Number of times a participant fixates on the media before fixating on an AOI for the first time.
3. First Fixation Duration: Duration of the first fixation on an AOI.
4. Fixation Duration: Duration of each individual fixation within an AOI.
5. Total Fixation Duration: Sum of the duration for all fixations within an AOI.
6. Fixation Count: Number of times a participant fixates on an AOI.
7. Visit Duration: Duration of each individual visit within an AOI.
8. Visit Count: Number of visits within an AOI.

for several milliseconds we have a fixation. In our study, eye movements were captured using a Tobii 1750 eye tracker integrated into a 17" TFT monitor with a 1280×1024 resolution. When activated, the eye tracker illuminated the user with two infrared projections that generated reflection patterns on the corneas of the eyes. A video camera gathered these reflection patterns along with the position of the user and, through digital image processing, the pupil locations were extracted at a rate of 50 Hz. The pupil positions were then mapped to gaze locations on the screen. For the gaze behaviour analysis we used the eye metrics listed in Table 4, which were extracted automatically using the Tobii Studio Statistics application. The metrics were calculated based on defined Areas of Interest (AOIs) and data selection time intervals. We defined three AOIs: title, body of article, and associated comments.

3.5 Participants

For our user study, we recruited 57 participants (female=28, male=29) through a campus-wide ad, whose age ranged from 18 to 47. The participants had to be free from vision-related conditions and able to read what was displayed on the computer screen without difficulty and without wearing glasses. Moreover, the participants had to have adequate knowledge of the English language. To avoid any adverse effects due to language-specific bias, we evaluated their English language fluency by showing during the tutorial a sample news article and discussing its content. The participation fee was a 20-Euro gift card for a popular commercial centre.

The participants were of mixed ethnicity, including Spanish, Czech, British, German, Italian, and other. The majority (50.8%) had a Master's degree or some college degree (29.8%). They were primarily pursuing further studies while working (54.3%), although there were a number of students (36.8%) and full-time employees (7%). Participants were all proficient with the English language (27.3% intermediate level, 64% advanced level, 8.7% native speakers).

3.6 Experimental Methodology

The experiment had a three-way, mixed design. The related measures independent variable was the comment ordering ("popular now", "newest", "most replied", "sentimentality high", "sentimentality low", "polarity (+)", "polarity (−)"). The unrelated measures independent variables were the article category ("crime and law", "entertainment and lifestyle", "science") and article interestingness (two levels: "interesting", "uninteresting"). The comment ordering was controlled by providing links to all seven types of orderings. By clicking each link, the participants could view the corresponding comments, ordered in the ways discussed in Section 3.1. The article interestingness was controlled by showing articles that were perceived as interesting or uninteresting. Interestingness was established by asking the participants to rank in the entry questionnaire (Section 3.3) six news titles per article genre (18 titles in total), and assign the most interesting title to the first position, the next most interesting title to the second position, and so on, leaving the least interesting for the last. The news title ranked first determined the interesting article, while the news title ranked last determined the uninteresting article. These top- and bottom- ranked news articles were used to manipulate the interestingness levels. The primary dependent variable was participants' news reading behaviour as determined by the application of eye-tracking. Other dependent variables were participants' affect state before and after each task, level of focused attention, as well as participants' opinion about the news article and associated comments.

Each participant was asked to complete two tasks: the first task involved reading an interesting news article and the second task reading an uninteresting news article, each of a different genre. The genre was determined by the meta-information provided by Yahoo! News while article interestingness was established using a questionnaire (see Section 3.3). The tasks were presented in the context of a short cover story, which asked the participants to read a news article and all of the seven different orderings of associated comments, and answer an online survey. To control the order effects, the tasks assignment was counterbalanced by using a Latin Squares design. The read order of the seven different comment orderings was randomised.

Table 5 outlines the flow of the news reading task, from its introduction to the payment. The study was performed as follows. From the outset, the participants sat in a quiet room, facing the computer they would use to perform the news reading task. The experimenter was situated on the other side of the desk, concealed by a divider, from where he monitored the procedure. All instructions where handed in printed form. The opening instructions discussed the purpose of the study, described the equipment, addressed confidentiality and privacy issues, and outlined the experimental procedure. Once participants had finished reading the information sheet, the key points of the procedure were revisited, and any remaining questions were answered. The session began once the participant understood what s/he had to do and had signed a consent form. Following that, participants were asked to complete the entry questionnaire and then watch a video tutorial, which walked them through the task and also

Table 5: Experimental procedure

Introduction	Introduction and consent form
	Entry questionnaire (news articles ranking)
	Video tutorial
	Task scenario
Task (repeat. twice)	Pre-task PANAS
	Eye-tracker calibration
	News reading task (article and associated comments)
	Post-task PANAS
	Focused attention self-report
	Interest and task ease questions
Final	Demographics questionnaire
	Payment and receipt

indicated body postures that could impair the eye tracker. The height of the chair was then adjusted and the eye tracker was calibrated.

Participants were presented with two web browser windows: the first window showed the news article while the second window indicated the steps to follow along with the questionnaire. Upon reading the news article or a related category of comments, the participants switched to the questionnaire to answer the relevant section and then resumed their task. Participants were asked to proceed with reading the article and the comments in their own time. After the end of the first task the participants were given the option to take a short break. The same procedure, including the eye tracker calibration, was repeated for the second task and for a different news article. A brief discussion between the experimenter and the participant followed the completion of the second task, to ensure that the participants had no further questions.

4 Results

We present the findings based on 114 tasks carried out by 57 individuals. We report the questionnaire data that refer to the news articles and associated comments, and those that refer to the eye tracking data. For our analysis we used several related and unrelated measures tests, including Mann-Whitney, Wilcoxon Signed-Rank test and the dependent means t-test for pair-wise comparisons, One-way Repeated measures ANOVA for three or more conditions, Chi-Square Test of Association for categorical data, and Spearman's rho. Participants responses to the 10-item PAS, 10-item NAS, and 9-item focused attention scale were summed to obtain the final scores. Results are reported at a statistical significance level of .05. To take an appropriate control of Type I errors in multiple pair-wise comparisons we applied the Bonferroni correction. The data ($N = 114$) were examined for missing values that were NMAR (not missing at random), which resulted in the exclusion of two participants' data from subsequent analysis.

4.1 Questionnaire Data

4.1.1 Article

We present participants' experience with the news articles and associated comments. To the "I found the news article interesting to read" question, participants' assessment was consistent with our experimental manipulation of article interestingness, reporting on average higher levels of interest for the "interesting" ($M = 3.79, SE = .13$) compared to the "uninteresting" condition ($M = 2.58, SE = .13$). To the "I was familiar with the content of the news article" statement participants were neutral in the "interesting" condition ($M = 3.03, SE = .14$) and less in agreement in the "uninteresting" condition ($M = 1.86, SE = .14$). Participants also found the news articles easy to understand ($M = 3.70, SE = .14$). To the "I typically read this type of news on the web" statement participants indicated higher agreement in the "interesting" condition ($M = 3.58, SE = .14$) than the "uninteresting" condition ($M = 2.11, SE = .13$). Finally, they reported higher agreement with the statement "'I wanted to find out more about the news story I read" when it involved interesting articles ($M = 3.33, SE = .148$) rather than uninteresting ($M = 2.04, SE = .14$).

When asked to indicate which ways they prefer to stay informed about current news, 53 participants reported reading the news online between "once or twice a week" to "one or more times a day" ($M = 4.45, SE = .15$), 32 participants reported watching the news on the television, 28 reading the printed press, and 16 listening to the radio news broadcast. On average, female participants spent 27 minutes reading the news online, while male participants spent 39 minutes. To the "How often do you leave comments in the news articles you read online?" question participants' responses varied between "never" to "rarely" ($M = 1.51, SE = .09$). When asked how often they browse other people's comments in news articles they read online, responses varied between "rarely" to "sometimes" ($M = 2.64, SE = .12$), which agrees with their view of how important comments are to the overall news reading experience ($M = 2.87, SE = .12$).

4.1.2 Comments

To evaluate the effect of the different comment orderings we asked our participants to report their agreement to the following statements: (i) the comments I read were interesting, (ii) the comments provided additional insight for the article, and (iii) I enjoyed reading these comments. Figures 4, 5, and 6 show the

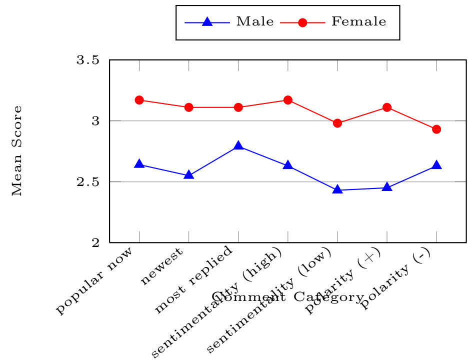

Figure 4: Reported interest across all comment orderings.

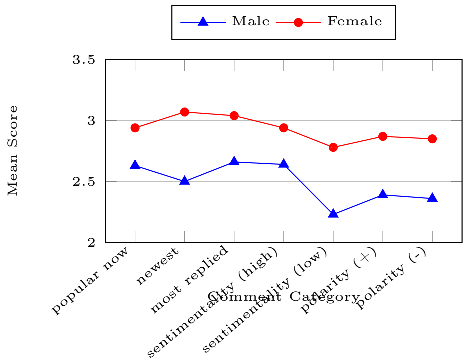

Figure 5: Reported level of insights across all comment orderings.

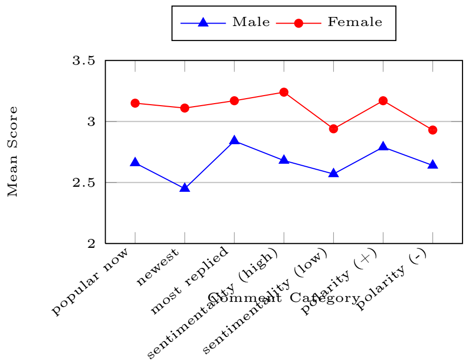

Figure 6: Reported enjoyment across all comment orderings.

average scores across all comment orderings and genders. In all cases, the scores reported by the female participants were significantly higher compared to those of male participants, $U = 514628, z = -9.799, p < .001, r = -0.20$, suggesting that they had became more "positively engaged" with the associated comments.

With respect to statement (i), the Friedman's ANOVA test indicated that male participants' scores changed significantly across the comment orderings, ($\chi^2(56) = 12.568, p < .05$), whereas the scores reported by the female participants did not differ significantly. The main message here is that female participants perceived as equally interesting the comments shown in different orderings. For male participants, the comment orderings with the highest average scores on the interestingness scale were the "most replied" and "sentimentality high".

For statement (ii), we observe a somewhat similar trend where only the scores reported by male participants are significantly different ($\chi^2(56) = 21.016, p <$

.005). Again, we notice that the most insightful comments were those listed in "most replied" and "sentimentality high". This is in alignment with our assumption that sentiment is essential for most social media because it typically centres on human interaction, which can be dull without opinions or emotions. Another thing to note is that, given the low scores of "polarity (+)" and "polarity (−)", the polarity of the expressed sentiment was not as important as the intensity.

Finally, responses to statement (iii) were found to be statistically different for the male participants only ($\chi^2(56) = 14.074, p < .05$). Figure 6 shows that the most replied and positive comments invoked stronger feelings of enjoyment, compared to other comments. In some cases, the new comment orderings ("sentimentality high", "sentimentality low", "polarity (+)", "polarity (−)") had a similar impact to the news reading experience as the default ones ("popular now", "newest", "most replied"), which did not apply any systematic control over the sentiment expressed in the content. This is a noteworthy finding because it points out the practical value of sentiment in promoting certain dimensions of engagement, such as interest and experienced affect.

Wilcoxon tests were used to follow up the findings for statements (i), (ii), and (iii). A Bonferroni correction was applied and so all effects were examined at a 0.002 level of significance. The tests did not reveal any significant difference between any of the pair-wise comparisons of comment orderings. However, given the very conservative approach of the Bonferroni correction, applying such a corrective factor poses the risk of a type II error and, as a result, reduces the statistical power of our analysis. Especially in the case of a large number of tests, this can result in critical values so small that no differences are possible.

4.2 Examining Interest via User Engagement Metrics

We report the experimental findings of our laboratory study, which demonstrates how user engagement is spanned across online news of varying interestingness. Given the multi-faceted nature of user engagement, we employ a set of subjective and objective measures, such as positive and negative affect, focused attention, and gaze behaviour.

4.2.1 Controlled and Perceived Interest

We recall that our experimental manipulation of article interestingness was based a pre-task ranking of news titles, performed by each participant individually. In addition, we asked our participants to report the degree of experienced interest in the news article by responding to the post-task question "Based on your experience with the news article you read, please indicate to what extend you agree or disagree with the following statement: I found the news article interesting to read". To validate our experimental manipulation of interestingness we triangulated the initial news title ranking with the reported scores to the post-task questionnaire. News articles that received scores between 4 and 5 were treated as interesting, while those that received scores between 1 and 2 were treated as uninteresting. The remaining were treated as neutral and were excluded from our analysis.

The Chi-Square test revealed a significant association between the article interestingness and the perceived interest ($\chi^2 = 29.52, p < .001$), with a strong positive relationship ($\phi = .518, p < .001$), which confirms the effect of our experimental manipulation. The follow-up analysis of the hypotheses is based on the levels of article interestingness as we manipulated it experimentally, with the exception of focused attention, where we referred to perceived interest for testing the significance of the observed differences.

4.2.2 Interest and Experienced Affect

The Wilcoxon Signed-Rank test was applied to determine the significance of the variance observed in the reported negative and positive emotions, before and after each task. We examined the difference between the pre-negative (preNAS) and post-negative (postNAS) affect scale, and pre-positive (prePAS) and post-positive (postPAS) affect scale, for both experimental conditions ("interesting", "uninteresting"). The results did not indicate any significant difference between preNAS and postNAS, suggesting that there was no interaction effect between the experienced negative emotions and the article interestingness. Positive affect, however, decreased significantly from pre- to post-task, for both "interesting" ($T = 18.5, p < .001, r = -.58$) and "uninteresting" ($T = 19.0, p < .001, r = -.58$) tasks, most likely due to the experienced fatigue. In both cases, the r values indicated a large effect. When comparing the postPAS for both conditions, we observed a statistically significant decrease from "interesting" to "uninteresting" ($T = 431.5, p < .05, r = -.19$), indicating that the effect of the "uninteresting" task on the intensity and range of participants' positive emotions was larger. No significant difference was observed in the postNAS. Responses to the "Overall, did you feel positive or negative while completing the news reading task?" question revealed a statistically significant difference by condition, with the "interesting" task being associated with positive affect and the "uninteresting" task with negative affect ($\chi^2 = 10.53, p < .05$). In summary, we observed an improvement on affect when the news content was interesting, which was characterised by more intense and varied positive emotions. Considering that affect is a main dimensions of user engagement, this finding highlights the adverse effects of news content when it fails to appeal to the reader's interests.

4.2.3 Interest and Focused Attention

The Wilcoxon Signed-Rank test was used to test the differences in the focused attention across the article interestingness levels, but did not reveal any significant effects ($T = 554, p = .152$). The same analysis was repeated by also grouping the participants according to "perceived interest" (as reported in the questionnaire). Given that the samples in this type of grouping were treated as independent, we opted for the Mann-Whitney test. Participants in the "interesting" condition ($Mdn = 47.46$) reported significantly higher levels of attention compared to participants in the "uninteresting" condition ($Mdn = 32.19$), $U = 502, p < .01, r = -.31$. However, there was no significant effect for article interestingness in the perceived time ($t(74) = -.604, p = .548$) spend on the news reading task, the actual time ($t(74) = -1.246, p = .217$), or their difference ($t(74) = -.797, p = .428$). This finding suggests that interest in the news content lead to higher levels of focused attention, which is central to the concept of engagement.

4.2.4 Interest and Gaze

The dependent means t-test was used to analyse whether the differences observed across the experimental conditions were statistically significant. We computed the t-value for all metrics shown in Table 4, and for all the AOIs (title, body of article, and comments). For the news title AOI the only significant effect was observed for Visit Duration. On average, participants spend significantly more time ($t(50) = 2.19, p < .05, r = .29$) browsing the titles of interesting articles ($M = 1.00, SE = .09$), than those of uninteresting articles ($M = .77, SE = .06$).

Regarding the body of article AOI, the t-test revealed that the male participants performed significantly less fixations outside of the article AOI ("Fixations Before"), when reading interesting news articles. Furthermore, the time until the first fixation ("Time to First Fixation") occurred in shorter time for interesting news articles. The results summarised in Table 6 indicate medium to large effect sizes of interestingness on gaze behaviour and attention. No significant differences were observed for the female participants.

The t-test was also used to analyse the eye movements in terms of the comments AOI, for all orderings of comments. There was a significant effect for the "Fixation Count" metric for the "newest" category ($t(51) = 2.35, p < .05, r = .31$). Participants performed more fixations in the interesting news articles ($M = 150.78, SE = 10.34$), as opposed to the uninteresting ones ($M = 123.20, SE = 7.51$). The t-test revealed a significant effect for "Visit Count", for the "most replied" ($t(51) = -2.22, p < .05, r = .29$) and "polarity ($-$)" ($t(51) = 2.15, p < .05, r = .28$) orderings. Also, participants looked at the "most replied" ($M = 16.25, SE = 1.48$) and "polarity ($-$)" ($M = 15.23, SE = 1.76$) orderings of comments more often when reading interesting articles, compared to uninteresting ($M = 12.69, SE = 1.18, M = 11.94, SE = 1.08$). These findings suggest that users are more likely to allocate time and attention reading other people's comments, and potentially be part of this online social exchange, when the comments are perceived as interesting compared to when they are not.

To analyse the observed intra-comments variability for all eye metrics and interestingness conditions we used the One-Way Repeated Measures ANOVA test. The mean and standard deviation scores are shown in Table 7. For "Total Fixation Duration" the Mauchly's test indicated that the assumption of sphericity had been violated for both "interesting" ($\chi^2(20) = 37.28, p < .001$) and "uninteresting" conditions ($\chi^2(20) = 35.87, p < .001$). Therefore, degrees of freedom

Table 6: t-test results from the analysis of eye metrics between the two experimental conditions (interesting versus uninteresting articles)

Eye metrics	Condition	M	t	r
(Male participants)				
Time to First Fixation	interesting	11.53	-3.41**	0.55
	uninteresting	64.46		
Fixations Before	interesting	8.55	-2.03*	0.37
	uninteresting	15.26		
First Fixation Duration	interesting	0.27	-0.26	-
	uninteresting	0.29		
Fixation Duration	interesting	0.38	1.51	-
	uninteresting	0.36		
Total Fixation Duration	interesting	155.58	1.53	-
	uninteresting	129.04		
Fixation Count	interesting	401.7	1.20	-
	uninteresting	346.92		
Visit Duration	interesting	6.42	-0.75	-
	uninteresting	6.99		
Visit Count	interesting	36.22	1.65	-
	uninteresting	28		
(Female participants)				
Time to First Fixation	interesting	40.31	-1.63	-
	uninteresting	102.75		
Fixations Before	interesting	12.08	-1.43	-
	uninteresting	30.68		
First Fixation Duration	interesting	0.36	-1.12	-
	uninteresting	0.47		
Fixation Duration	interesting	0.42	0.63	-
	uninteresting	0.40		
Total Fixation Duration	interesting	236.82	1.69	-
	uninteresting	167.77		
Fixation Count	interesting	566.6	1.71	-
	uninteresting	402.16		
Visit Duration	interesting	5.13	0.94	-
	uninteresting	4.18		
Visit Count	interesting	64.24	1.67	-
	uninteresting	43.84		

*. Statistical significance at the .05 level (2-tailed).
**. Statistical significance at the .01 level (2-tailed).

Table 7: Descriptive statistics of the eye metrics for the two experimental conditions ("interesting" and "uninteresting" news articles)

		interesting		uninteresting	
	Comment ordering	*M*	*SD*	*M*	*SD*
Total Fix. Dur.	most replied	52.95	32.76	55.09	25.70
	newest	59.53	30.37	49.98	27.02
	polarity (−)	43.36	24.44	41.53	15.29
	polarity (+)	51.41	31.38	44.97	22.12
	popular now	59.25	29.53	52.34	33.01
	sentimentality high	52.14	31.45	55.01	24.01
	sentimentality low	45.12	23.20	40.05	22.81
Fix. Count	most replied	134.65	75.40	142.40	62.91
	newest	150.78	74.57	123.20	54.18
	polarity (−)	113.78	56.26	107.62	36.23
	polarity (+)	127.39	66.67	115.07	50.79
	popular now	148.52	65.29	128.48	64.03
	sentimentality high	130.20	68.80	140.68	59.46
	sentimentality low	115.86	51.72	99.62	51.93
Visit Dur.	most replied	-	-	4.66	3.49
	newest	-	-	5.32	3.78
	polarity (−)	-	-	4.73	3.27
	polarity (+)	-	-	4.26	2.34
	popular now	-	-	4.64	3.85
	sentimentality high	-	-	5.75	5.14
	sentimentality low	-	-	3.79	2.06

were corrected using Huynh-Feldt estimates of sphericity ($\epsilon = .91$ for "interesting", $\epsilon = .90$ for "uninteresting"). The results show that the total fixation duration associated with the comments AOI was significantly affected by the category of comments, for both "interesting" ($F(5.47, 278.92) = 3.72, p < .01$) and "uninteresting" ($F(5.43, 277.36) = 4.67, p < .001$) news articles. The One-Way Repeated Measures ANOVA also determined that "Fixation Count" ($F(6, 306) = 3.90, p < .001$) and "Visit Duration" ($F(4.79, 244.30) = 2.51, p < .05$) differed statistically significantly between comment categories, for both experimental conditions. For "Visit Duration" the Mauchly's test indicated that the assumption of sphericity was violated ($\chi^2(20) = 61.20, p < .001$). The degrees of freedom were corrected using Huynh-Feldt estimates of sphericity ($\epsilon = .79$). We omit reporting further results from post-hoc tests for brevity.

Overall, the "Total Fixation Duration" scores suggest that participants in the "interesting" condition remained fixated for longer periods when reading the "newest" and "popular now" comments. However, participants in the "uninteresting" condition fixated for longer periods when reading the "most replied" and "sentimentality high" comments. The same trend is observed for "Fixation Count" and "Visit Duration", reflecting users' preference of certain comment orderings over others, depending on the experimental condition. Based on these findings, we can conclude that a connection exists between interestingness and gaze behaviour.

4.3 Correlation Analysis of all Factors

We report the results of the correlation analysis performed across all factors, where we examined the effect of expressed sentiment in news articles on user attention, interest, affect, and gaze. The importance of this analysis is the finding of significant associations between features like sentimentality and polarity, which are scalable and can be extracted automatically from the content, as well as offline behaviour measures like affect or gaze, which have more discriminative power but are also more laborious to collect. Such correlations are useful because they indicate a predictive relationship that can be exploited in practice by leveraging the scale- and scope- specific limitations that each type of engagement measure introduces. When our data violated the normality assumption we opted for the Spearman's rank correlation coefficient test.

4.3.1 Affect, Focused Attention and Sentiment

We run a correlation analysis to determine the association between focused attention, postPAS, postNAS, the change from prePAS to postPAS, and the change from preNAS to postNAS, with the news content sentimentality and polarity. The sentimentality and polarity scores were further blocked into those relevant to the news titles and those relevant to the body of articles. Out of the many pair-wise comparisons, only those between the change in positive affect, sentimentality, and polarity were found to be statistically significant. The change in positive affect had a positive relationship with the sentimentality of the title ($r_s = .28, p < .01$) and the body of the article ($r_s = .25, p < .01$), and a negative relationship with the polarity of the title ($r_s = -.41, p < .001$) and the body of the article ($r_s = -.29, p < .01$). This suggests that positive affect increased as the title and body of the news article became more sentimental and more negative.

4.3.2 Interest, Enjoyment, Curiosity and Sentiment

We examined the relationship between the sentimentality and polarity of news content with interest, experienced enjoyment, and wanting to know more about the news. Spearman's rank correlation coefficient test indicated a significant positive correlation between perceived interest and title sentimentality ($r_s = .21, p < .05$), and a highly significant negative correlation with title polarity ($r_s = -.35, p < .001$) and body of article polarity ($r_s = -.36, p < .001$). Similarly, experienced enjoyment (as a result of reading the news article) was also found negatively associated with article polarity ($r_s = -.22, p < .05$). Finally, curiosity to know more about the article was found to be negatively correlated with title polarity ($r_s = -.31, p < .001$) and article polarity ($r_s = .32, p < .001$).

Table 8: Summary of inter-correlations of eye metrics and news content sentimentality and polarity scales

	Title		Article	
Eye metrics	Sentimentality	Polarity	Sentimentality	Polarity
Time to First Fixation	.06	.07	.12	-.18
Fixations Before	.08	.12	.09	-.15
First Fixation Duration	-.08	.02	.05	-.03
Fixation Duration	-.12	.10	-.15	-.03
Total Fixation Duration	-.21*	.19*	-.09	.13
Fixation Count	-.20*	.18	-.03	.15
Visit Duration	-.21*	.16	.02	-.04
Visit Count	-.11	.14	-.03	.06

*. Correlation is significant at the .05 level (2-tailed).

Here we observe that user engagement was enticed by news content with strong sentimentality and negative connotations.

4.3.3 Gaze Behaviour and Sentiment

The Spearman's rank correlation coefficient test was used to measure the association between gaze behaviour and the sentimentality and polarity of news content. The gaze metrics were examined separately for the news title AOI and the body of article AOI. The inter-correlations shown in Table 8 indicate a negative relationship between "Total Fixation Duration" and title sentimentality ($r_s = -.21, p < .05$), as well as a positive relationship with title polarity ($r_s = .19, p < .05$). Similarly, "Fixation Count" ($r_s = -.20, p < .05$) and "Visit Duration" ($r_s = -.21, p < .05$) were found to be negatively correlated with the title sentimentality. This finding suggests that the participants fixated more often, and for longer time, when reading news titles of weak sentiment and positive tone.

4.3.4 Affect, Focused Attention and Gaze Behaviour

A correlation analysis was applied to examine the association between the eye metrics and postPAS, postNAS, the change from prePAS to postPAS, the change from preNAS to postNAS, and focused attention. The analysis was performed separately for the news title and the body of article AOI's. The results are presented in Table 9, showing several positive and negative associations. For example, "Total Fixation Duration" ($r_s = .26, p < .01$), "Fixation Count" ($r_s = .29, p < .01$), and "Visit Count" ($r_s = .29, p < .01$) appear to be positively correlated with postPAS, revealing that the participants fixated more often and for longer time at news titles when being in a state of positive affect. A similar correlation is also observed for the metric "Visit Count" ($r_s = .21, p < .05$), in regards to the body of article AOI. Finally, for the same AOI, we can conclude that participants performed their first fixation faster ($r_s = .26, p < .01$) when

Table 9: Inter-correlations of eye metrics with affect and focused attention

	FA	postPAS	postNAS	pre/postPAS	pre/postNAS
News Title AOI					
Time to First Fixation	.03	-.06	.02	-.04	-.06
Fixations Before	.05	-.07	.02	-.09	-.06
First Fixation Duration	-.07	-.08	.06	-.12	-.15
Fixation Duration	-.01	-.02	.01	-.07	-.17
Total Fixation Duration	-.10	.26**	.04	.02	-.004
Fixation Count	-.12	.29**	.04	.03	.04
Visit Duration	-.11	.05	-.006	-.10	-.09
Visit Count	-.04	.29**	.06	.09	.06
News Article AOI					
Time to First Fixation	.10	-.03	.04	.26**	-.09
Fixations Before	.11	-.02	.04	.29**	-.06
First Fixation Duration	.08	-.14	.0009	-.07	-.13
Fixation Duration	.04	.12	-.03	-.14	-.04
Total Fixation Duration	.02	.11	.17	-.25**	-.03
Fixation Count	.01	.07	.17	-.21*	-.002
Visit Duration	.04	.21*	-.06	-.02	.10
Visit Count	.03	-.06	.14	-.18	-.11

*. Correlation is significant at the .05 level (2-tailed).
**. Correlation is significant at the .01 level (2-tailed).

positive affect increased from pre- to post- task, whereas the "Total Fixation Duration" ($r_s = -.25, p < .01$) and the "Fixation Count" ($r_s = .21, p < .05$) were reduced.

4.3.5 Article Interest and Associated Comments

In addition, we analysed the relationship between article interestingness and the associated comments, for the following statements: (i) "the comments I read were interesting", (ii), "the comments provided additional insight for the article", (iii) "the comments improved my news reading experience", (iv) "I enjoyed reading these comments", and (v) "I found the comments useful in understanding better the article". Since article interestingness is a dichotomous variable, we computed the biserial correlation coefficient (r_b). The biserial correlation analysis indicated a significant association between article interestingness and interest in the comments ($r_b = .18, p < .001$), as well as article interestingness and level of enjoyment in reading the comments ($r_b = .11, p < .01$). For all other pair-wise comparisons, the results were not found to be statistically significant. In short, interest in news content was associated with the reading experience facilitated by the reader's comments, highlighting the importance of user-generated content in social interaction.

5 Discussion

In this paper we investigated the connection that exists between sentimentality and polarity of news content and user engagement, and attempted to establish the extent to which certain aspects of the interaction influence engagement, while observing the fundamental role of user interest. Our methodological approach involved an analysis of 13,319 news articles taken from Yahoo! News US and a controlled study, which allowed us to assess several facets of user engagement. Previous studies have examined the relationship between information propagation and content properties like sentimentality and polarity, or user-supplied comments. To the best of our knowledge, none of the previous work has tried to quantify the level of interaction between the user and the content in such a way, for the online news domain.

We began by showing that news articles exhibit a considerable variation in terms of the sentimentality and polarity of their content, and examined these differences with respect to the news source they are obtained from, the genre, and the time they were published on the Web. Only the news source and the genre had a significant effect on the sentiment metrics of news content, with genre being the most prominent feature. Our news collection included 14 different genres of news with distinctive sentimentality and polarity patterns. For example, law-, health-, or politics- related news were characterised by highly sentimental, negative content, unlike the more neutral technology and science genres, although the general trend observed in the news press can be described as negative compared to other types of media. This finding reinforced our initial assumption that news genres vary significantly on the sentimentality and polarity of their content and, thus, are more likely to have a diverse effect on how much users engage with the news stories.

Having established this variation, we used a smaller subset of our news collection to demonstrate in a laboratory setting how user engagement differs for news articles of varying interestingness. Given the multi-faceted nature of user engagement, we employed a set of subjective and objective measures, such as positive and negative affect, focused attention, and gaze behaviour. As mentioned previously, a significant decrease was observed in PAS when comparing interesting and uninteresting news articles. This finding was further supported by participants' responses to the "Did you feel positive or negative while completing the news reading task?" question, which also revealed a significant difference by condition, with the interesting news being associated with positive affect and the uninteresting with negative affect. The article interestingness had no impact on the negative affect scale and, therefore, the contribution of NAS here was limited. We speculate that the decrease from pre- to post- PAS and the absence of any effect on NAS was a result of the way the PANAS was administered before and after the news reading task. However, introducing the PANAS survey separately for the article and the comments would have induced a long, repetitive experience that could have biased our findings. Despite this limitation, we observed a clear effect of the interestingness variable, suggesting that *affect improves when the news content is interesting compared to when it is* *not*. This finding highlights the emotional and cognitive connection that exists between the user and the news content, and also the adverse effects of news content when it fails to appeal to the reader's interests.

Our examination of the interaction between focused attention and interestingness did not reveal any significant effect. The same analysis was repeated after blocking our participants according to the reported interest (which was found highly correlated to the interestingness levels we introduced) and showed a significant difference. *Participants who read interesting articles exhibited much higher levels of focused attention compared to participants who read uninteresting articles.* Since focused attention is central to the concept of engagement, this finding corroborates further the importance of the resource (in our case the news content) on the degree of engagement induced by it. Other measures, like perceived time spend on the news reading task or actual time, did not vary significantly. Partly, the role of estimated time (and its comparison to actual time) does not seem to be as important, especially when the tasks are not long enough to induce some difference. Similar findings were reported in [28].

We demonstrated an approach to quantifying the effect of user interest on attention and focus, both important dimensions of user engagement, by examining gaze behaviour. We observed how eye metrics like "Total Fixation Duration", "Fixation Count", and "Visit Count", differed across online news of varying interestingness and their associated comments. Our analysis indicated that participants spend significantly more time browsing the titles of interesting articles, performed more gaze visits, as well as quicker to occur, more frequent, and prolonged fixations. These quantitative results are in accordance with the qualitative findings from the PANAS and Focused Attention tests, and confirm the fundamental role of user interest in promoting engagement, the mid-point between the creation of online content and the wide circulation of the Web. Although this may not be such a novel contribution [28, 30], it is still a methodologically sound replication of previous findings, both in terms of measurement and analysis.

Our exploratory analysis of the interactions between all the factors accounted for in this study indicated several significant correlations. Foremost, we found a significant association between positive affect and news sentiment. Positive affect showed a positive correlation with title and article sentimentality, and a negative correlation with title and article polarity. The same trend was also observed with reported interest, enjoyment, and curiosity to know more. Perceived interest was found to be negatively correlated with title sentimentality and polarity, and article polarity. Similarly, enjoyment in reading the news was found negatively correlated with article polarity, and curiosity to know more was negatively correlated with the title and the article polarity. The message to be taken is that *interest, enjoyment, and the desire to find out more about online news is enticed when strong sentiment and negative connotations are present in the content.* This makes sense; news headlines are written to induce emotional reactions and attract attention. Based on the initial impression of the title, the user will decide whether to proceed with reading the article or ignore it and focus on other news. Also, given the variation of the sentimentality and polarity of news content, this finding reveals an important predictive relationship that could be exploited further by leveraging measures like gaze or affect, which are difficult to collect, with more scalable features that derive from the content, i.e., sentimentality and polarity.

We also examined the association of gaze behaviour and sentiment. Contrary to previous findings, our gaze analysis revealed that longer and more frequent fixations occurred when users were viewing news titles characterised by low sentimentality and positive polarity. Gaze was also found to be correlated with positive affect. This is another noteworthy finding in terms of predicting user engagement, since it suggests that gaze behaviour can act as a proxy for sentiment, and vice versa. In other words, it is measurable and to a point can be anticipated. In terms of practical implications for news providers, it is important to understand what happens when the user is exposed to new information. It is during this critical attention process that the user decides whether the content is worth reading, and without that mid-layer of interaction the effective reach of information is constrained.

Moreover, we observed a relationship between interest in a news article and the reading experience facilitated by user comments, left in response to the news article. Our qualitative analysis sheds further light on the effect of different ranking methods (which systematically exploit the sentiment expressed in user comments) on interest, experienced enjoyment, and insights. Among our main findings is a gender-specific bias that characterises the degree of experienced engagement, as well as a preference over highly sentimental or popular comments. In the majority of cases, female participants were significantly more interested in the user comments and reported higher levels of enjoyment across all comment orderings. Contrary to our male participants, the existence of different categories did not affect their view of the comments nor resulted in any specific preferences. Male participants appeared to be less engaged with the user comments, but also favoured certain comment orderings over others. This connection between gender and user-generated content is an interesting finding that warrants further investigation, to determine the extent to which such bias defines the reading experience and how it can inform the design of personalisation techniques or improve content adaptation.

High volume of user-generated content, in our case comments posted with news articles, is an indication of engaged users. Users interested in the content they read usually take an "active" online role and comment on it, which also suggests that they take into consideration other people's comments. This behaviour is in accordance to our social nature and our innate tendency to define ourselves in relation to others, in a social context that both frames and directs our actions and experiences. With the growing interactivity of news, and the importance of user-generated content, the user-driven discussions anchored around online news stories are of paramount importance to news providers that wish to stimulate public interest and promote engagement, both in terms of further interaction with the news site (spending more time on the news site, posting comments). Given that *certain orderings of comments are more preferable over others, depending on the interestingness of the article and the gender of the* *user*, news portals can determine how to benefit from user-driven commenting to promote higher levels of engagement.

Finally, our results *highlighted the importance of interesting content to keep readers engaged.* However we should bear in mind that what gets published by the news media depends on numerous factors, an important one being the newsworthiness of a story, but also factors such as space constraints, timeliness, and how *close* a story is to its readers, for instance, geographically, culturally, or simply in terms of general interests (e.g. sports, international news, gossip columns) [16]. Since it is impossible to report everything, selectivity in what gets published is inevitable. Nonetheless, reputable news media are expected to be objective in which stories they decide report; their role is to inform people about what is happening. It is their task to find the right trade-off between interesting and non-interesting, but nonetheless informative, news-worthy content. This is important for both a fair and knowledgeable society, as media indeed can influence people's perception of the world including things closer to them [29].

6 Conclusions

This work examined the relationship between sentiment metrics that derive from the content and offline behaviour measures of engagement, for the online news domain. Our goal was to determine which aspects of user interaction influence engagement the most in the online news domain, while observing the fundamental role of user interest. In addition, we have demonstrated how real-life data, together with appropriate metrics, can leverage otherwise difficult to collect engagement measures and help quantify the level of interaction between the user and the news content. Despite the cost and effort for conducting such a study, we were able to perform a thorough exploration of diverse approaches to assessing user engagement, as well as account for the affective and cognitive aspects. The main message that stems from this study is that user engagement can be anticipated to some degree if we account for the factors that inspire human curiosity and drive attention, such as the sentimentality and polarity of content.

Our work also comes with certain limitations. One of them was the relatively small sample of the population we studied, which, to some extent, has affected our ability to generalise our findings to the population as a whole. However, this is a very common caveat in user studies and, in our case, a necessary trade-off, given the controlled and time-demanding nature of our experiment. An additional limitation was the moderate size of our news dataset, which was obtained from a single news provider. Given a much larger news article collection, the sentimentality and polarity scores we computed would be more accurate and exhibit less variation. Additionally, we used a lexicon-based sentiment analysis tool for computing the sentiment features that may not be completely accurate.

We have several avenues for our future work. First, we will employ a larger and more diverse news collection, to achieve a higher degree of ecological validity. Moreover, we will employ additional sentiment analysis tools, to allow for comparability of results with respect to the news domain. We will also investigate the extent to which gender-bias applies on our current findings within a larger sample of the population. Finally, we will examine cross-platform applications (e.g., social media platforms and blogs) of our approach to assessing engagement, where we will likely observe very different results for user-generated content and engagement, framed in a social interaction context.

7 Acknowledgment

We would like to thank our participants and all the people who volunteered for the evaluation of the news article collection. We would also like to thank the reviewers for their detailed and helpful comments.

References

[1] Abbasi, A., Chen, H., and Salem, A. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems 26*, 3 (2008), 12:1–12:34.

[2] Attfield, S., Kazai, G., Lalmas, M., and Piwowarski, B. Towards a science of user engagement (position paper). In *WSDM Workshop on User Modelling for Web Applications* (New York, NY, USA, February 2011), WSDM '11, ACM.

[3] Bai, X. Predicting consumer sentiments from online text. *Decision Support Systems 50* (2011), 732–742.

[4] Bandari, R., Sitaram, A., and Huberman, Bernardo, A. The pulse of news in social media: Forecasting popularity. In *Proceedings of the Sixth International Conference on Weblogs and Social Media* (June 2012), ICWSM '12, The AAAI Press.

[5] Bautin, M., Vijayarenu, L., and Skiena, S. International sentiment analysis for news and blogs. In *Proceedings of the Second International Conference on Weblogs and Social Media* (March 2008), ICWSM '08, The AAAI Press.

[6] Bautin, M., Ward, C. B., Patil, A., and Skiena, S. Access: news and blog analysis for the social sciences. In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), WWW '10, ACM, pp. 1229–1232.

[7] Beineke, P., Hastie, T., Manning, C., and Vaithyanathan, S. Exploring sentiment summarization. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications* (2004), pp. 1–4.

[8] Berger, J. A., and Milkman, K. L. Social transmission, emotion, and the virality of online content. *Social Science Research Network Working Paper Series* (Dec. 2009).

[9] Buscher, G., Biedert, R., Heinesch, D., and Dengel, A. Eye tracking analysis of preferred reading regions on the screen. In *Proceedings of the 28th of the International Conference extended abstracts on Human Factors in Computing S* (New York, NY, USA, 2010), CHI EA '10, ACM, pp. 3307–3312.

[10] Cohen, J. *Statistical power analysis for the behavioural sciences (2nd ed.).* New York: Academic Press, 1988.

[11] Cohen, J. A power primer. *Psychological Bulletin 112*, 1 (1992), 155–159.

[12] Counts, S., and Fisher, K. Taking it all in? visual attention in microblog consumption. In *Proceedings of the Fifth International Conference on Weblogs and Social Media* (July 2011), ICWSM '11, The AAAI Press.

[13] Dave, K., Lawrence, S., and Pennock, D. M. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web* (2003), pp. 519–528.

[14] Devitt, A., and Ahmad, K. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (Prague, Czech Republic, June 2007), Association for Computational Linguistics, pp. 984–991.

[15] Fischer, M. H. An investigation of attention allocation during sequential eye movement tasks. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology 52*, 3 (1999), 649–677.

[16] Galtung, J., and Ruge, M. H. The structure of foreign news. *Journal of Peace Research 2*, 1 (1965), 64–91.

[17] Gerani, S., Carman, M. J., and Crestani, F. Investigating learning approaches for blog post opinion retrieval. In *Proceedings of the 31th European Conference on Information Retrieval* (2009), pp. 313–324.

[18] Godbole, N., Srinivasaiah, M., and Skiena, S. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media* (March 2007), ICWSM '07, The AAAI Press.

[19] Gregory, M. L., Chinchor, N., Whitney, P., Carter, R., Hetzler, E., and Turner, A. User-directed sentiment analysis: visualizing the affective content of documents. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text* (2006), pp. 23–30.

[20] Gwizdka, J., and Lopatovska, I. The role of subjective factors in the information search process. *Journal of the American Society for Information Science and Technology 60*, 12 (2009), 2452 – 2464.

[21] Henderson, J. M. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences 7*, 11 (November 2003), 498–504.

[22] Hendricks, M. New naa study shows consumers are highly engaged with newspaper media content across platforms, 2012.

[23] Kahneman, D. *Attention and Effort (Experimental Psychology).* Prentice Hall, 1973.

[24] Kucuktunc, O., Cambazoglu, B. B., Weber, I., and Ferhatosmanoglu, H. A large-scale sentiment analysis for yahoo! answers. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2012), WSDM '12, ACM, pp. 633–642.

[25] Lerman, K., Blair-Goldensohn, S., and McDonald, R. Sentiment summarization: evaluating and learning user preferences. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (2009), pp. 514–522.

[26] Lopatovska, I. Searching for good mood: examining relationships between search task and mood. *Proceedings of the American Society for Information Science and Technology 46*, 1 (2009), 1–13.

[27] Manduchi, A., and Picard, R. Circulations, revenues, and profits in a newspaper market with fixed advertising costs. *Journal of Media Economics 22*, 4 (2009), 211–238.

[28] McCay-Peet, L., Lalmas, M., and Navalpakkam, V. On saliency, affect and focused attention. In *ACM SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, May 2012), CHI '12, ACM, pp. 541–550.

[29] McCombs, M., and Shaw, D. The agenda-setting function of mass media. *Public Opinion Quarterly 36*, 2 (1972), 176–187.

[30] O'Brien, H. L. Exploring user engagement in online news interactions. *Proceedings of the American Society for Information Science and Technology 48*, 1 (2011), 1–10.

[31] O'Brien, H. L., and Toms, E. G. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology 59*, 6 (2008), 938–955.

[32] O'Brien, H. L., and Toms, E. G. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology 61*, 1 (January 2010), 50–69.

[33] O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media* (May 2010), ICWSM '10, The AAAI Press.

[34] Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing* (2002), pp. 79–86.

[35] Pannasch, S., Schulz, J., and Velichkovsky, B. M. On the control of visual fixation durations in free viewing of complex images. *Attention, Perception, & Psychophysics 73*, 4 (2011), 1120–1132.

[36] Rayner, K. Eye movements and visual cognition: Scene perception and reading. In *Springer series in neuropsychology* (1992), K. Rayner, Ed., Springer-Verlag, pp. 46–65.

[37] Remington, R. W. Attention and saccadic eye movements. *Journal of Experimental Psychology: Human Perception and Performance 6*, 4 (1980), 726–744.

[38] Shepherd, M., Findlay, J. M., and Hockey, R. J. The relationship between eye movements and spatial attention. *Quarterly Journal of Experimental Psychology 38*, 3 (August 1986), 475–491.

[39] Szabo, G., and Huberman, B. A. Predicting the popularity of online content. *Communications of the ACM 53*, 8 (2010), 80–88.

[40] Tatar, A., Leguay, J., Antoniadis, P., Limbourg, A., de Amorim, M. D., and Fdida, S. Predicting the popularity of online articles based on user comments. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics* (New York, NY, USA, 2011), no. 67 in WIMS '11, ACM, pp. 1–8.

[41] Thelwall, M., Buckley, K., and Paltoglou, G. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology 62*, 2 (Feb. 2011), 406–418.

[42] Thelwall, M., Buckley, K., and Paltoglou, G. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol. 63*, 1 (2012), 163–173.

[43] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology 61*, 12 (August 2010), 2544–2558.

[44] Thelwall, M., Buckley, K., Paltoglou, G., Skowron, M., Garcia, D., Gobron, S., Ahn, J., Kappas, A., Küster, D., and Holyst, J. Damping sentiment analysis in online communication: Discussions, monologs and dialogs. In *Computational Linguistics and Intelligent Text Processing*, vol. 7817 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 1–12.

[45] Tsagkias, M., Weerkamp, W., and de Rijke, M. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (New York, NY, USA, November 2009), WIMS '11, ACM, pp. 1765–1768.

[46] Tsagkias, M., Weerkamp, W., and de Rijke, M. News comments: Exploring, modeling, and online prediction. In *Proceedings of the 32nd European conference on Advances in Information Retrieval* (Berlin, Heidelberg, 2010), ECIR '10, Springer-Verlag, pp. 191–203.

[47] Turney, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (2002), pp. 417–424.

[48] Vural, A. G., Cambazoglu, B. B., Senkul, P., and Tokgoz, Z. A framework for sentiment analysis in turkish: Application to polarity detection of movie reviews in turkish. In *Computer and Information Sciences III*, E. Gelenbe and R. Lent, Eds. Springer London, 2013, pp. 437–445.

[49] Wang, F.-Y., Carley, K. M., Zeng, D., and Mao, W. Social computing: From social informatics to social intelligence. *Intelligent Systems, IEEE 22*, 2 (March 2007), 79–83.

[50] Ward, C. B., Bautin, M., and Skiena, S. Identifying differences in news coverage between cultural/ethnic groups. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 03* (Washington, DC, USA, 2009), WI-IAT '09, IEEE Computer Society, pp. 511–514.

[51] Watson, D., Clark, L., and Tellegen, A. Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology 54*, 6 (1988), 1063–1070.

[52] Weber, I., Ukkonen, A., and Gionis, A. Answers, not links: extracting tips from yahoo! answers to address how-to web queries. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2012), WSDM '12, ACM, pp. 613–622.

[53] Wu, S., Tan, C., Kleinberg, J. M., and Macy, M. W. Does bad news go away faster? In *Proceedings of the Fifth International Conference on Weblogs and Social Media* (July 2011), ICWSM '11, The AAAI Press.

[54] Yi, J., Nasukawa, T., Bunescu, R., and Niblack, W. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining* (2003), pp. 427–434.

[55] Zhang, W., and Skiena, S. Trading strategies to exploit blog and news sentiment. In *Proceedings of the Fourth International Conference on Weblogs and Social Media* (May 2010), ICWSM '10, The AAAI Press.

[56] Zhang, W., Yu, C., and Meng, W. Opinion retrieval from blogs. In *Proceedings of the 16th ACM International Conference on Information and Knowledge Management* (2007), pp. 831–840.