

## DELIVERABLE SUBMISSION SHEET

**To:** Susan Fraser *(Project Officer)*  
EUROPEAN COMMISSION  
Directorate-General Information Society and Media  
EUFO 1165A  
L-2920 Luxembourg

**From:**  
Project acronym: PHEME Project number: 611233  
Project manager: Kalina Bontcheva  
Project coordinator The University of Sheffield (USFD)

### The following deliverable:

Deliverable title: Healthcare Application Prototype and User Evaluation Results  
Deliverable number: D7.3  
Deliverable date: 28 February 2017  
Partners responsible: King's College London (KCL)  
Status:  Public  Restricted  Confidential

is now complete.  It is available for your inspection.  
 Relevant descriptive documents are attached.

### The deliverable is:

- a document
- a Website (URL: .....)
- software (.....)
- an event
- other (.....)

Sent to Project Officer: <a href="mailto:Susan.Fraser@ec.europa.eu">Susan.Fraser@ec.europa.eu</a>	Sent to functional mail box: CNECT-ICT-611233 @ec.europa.eu	On date: 28 February 2017
--	--	------------------------------

**FP7-ICT Strategic Targeted Research Project PHEME (No. 611233)**

Computing Veracity Across Media, Languages, and Social Networks



**D7.3 – Application and Evaluation Results**

---

Anna Kolliakou (King’s College London)  
Arno Scharl (MODUL University Vienna)  
Rob Stewart (King’s College London)

<b>PHEME Contributors</b>	<b>BRC Contributors</b>	<b>External Contributors</b>
Arkaitz Zubiaga	George Gkotsis	Walter Rafelsberger
Kalina Bontcheva	Dave Chandran	Diana Maynard
Maria Liakata		Bo Wang
Michael Ball		

This deliverable reports on the algorithms developed and further progress completed as part of four WP7 demonstration studies on medication, legal highs, mental health stigma and self-harm and suicide. We also describe the deployment of the interactive dashboard with details of its quantitative and qualitative evaluation. We conclude with a series of recommendations to improve the dashboard functionality and support the user-centred qualities of the interface.

**Keywords:** tweets, annotation, legal highs, medication, self-harm, suicide, mental health stigma, dashboard, evaluation, usability

Nature: **Report**

Contractual date of delivery: **28.02.17**

Reviewed By: **Kalina Bontcheva, Thierry Declerck and Rob Procter**

Web links: [www.pHEME.eu](http://www.pHEME.eu)

Dissemination: **P**

Actual date of delivery: **28.02.17**

## D7.3 Application and Evaluation Results

### **PHEME Consortium**

This document is part of the PHEME research project (No. 611233), partially funded by the FP7-ICT Programme.

#### **University of Sheffield**

Department of Computer Science  
Regent Court, 211 Portobello St  
Sheffield S1 4DP, UK  
Tel: +44 114 222 1930  
Fax: +44 114 222 1810  
Contact person: KalinaBontcheva  
E-mail: [K.Bontcheva@dcs.shef.ac.uk](mailto:K.Bontcheva@dcs.shef.ac.uk)

#### **Universitaet des Saarlandes**

Computer Linguistics  
Postfach 15 11 50  
D-66041 Saarbrücken  
Germany  
Contact person: Thierry Declerck  
E-mail: [declerck@dfki.de](mailto:declerck@dfki.de)

#### **MODUL University Vienna GMBH**

Am Kahlenberg1  
1190 Wien  
Austria  
Contact person: Arno Scharl  
E-mail: [scharl@modul.ac.at](mailto:scharl@modul.ac.at)

#### **Ontotext AD**

Polygraphia Office Center fl.4,  
47A Tsarigradsko Shosse,  
Sofia 1504, Bulgaria  
Contact person: Georgi Georgiev  
E-mail: [georgiev@ontotext.com](mailto:georgiev@ontotext.com)

#### **ATOS Spain SA**

Calle de Albarracin 25  
28037 Madrid  
Spain  
Contact person: TomásPariente Lobo  
E-mail: [tomas.parientalobo@atos.net](mailto:tomas.parientalobo@atos.net)

#### **King's College London**

Strand  
WC2R 2LS London  
United Kingdom  
Contact person: Robert Stewart  
E-mail: [robert.stewart@kcl.ac.uk](mailto:robert.stewart@kcl.ac.uk)

#### **iHub Ltd.**

NGONG, Road Bishop Magua Building  
4th floor  
00200 Nairobi  
Kenya  
Contact person: Rob Baker  
E-mail: [robbaker@ushahidi.com](mailto:robbaker@ushahidi.com)

#### **SwissInfo.ch**

Giacomettistrasse 3  
3000 Bern  
Switzerland  
Contact person: Peter Schibli  
E-mail

#### **The University of Warwick**

Kirby Corner Road  
University House  
CV4 8UW Coventry  
United Kingdom  
Contact person: Rob Procter  
E-mail: [Rob.Procter@warwick.ac.uk](mailto:Rob.Procter@warwick.ac.uk)

## D7.3 Application and Evaluation Results

### **Executive Summary**

In the first part of the deliverable, we provide a summary of the algorithms developed for each of the 4 demonstration studies, their integration into the PHEME pipeline and dashboard and progress with our combined analysis of data from social media and the clinical record.

In the second part of the report, we describe the on-site usability test of the dashboard at the Maudsley NIHR Biomedical Research Centre in January 2017. The purpose of the testing was to assess the usability of the interface design, information flow and architecture as well as the functionality of the algorithms developed in the demonstration studies.

Six attendees completed a set of real-life tasks and provided their feedback through a series of open-ended questions and a 10-item System Usability Scale (SUS). The sessions typically lasted for an hour and were led by a task administrator.

Overall, the participants found the website to be a useful, robust and comprehensive tool and the majority completed most of the tasks with ease. However, the evaluation identified a few areas of improvement, typically in activating searches and locating menus.

This deliverable contains the task completion success rates, time to task completion, participant feedback and recommendations for improvement. A copy of the evaluation tasks and other material used in the sessions are included in the Appendices.

## Contents

<b>Executive Summary</b> .....	<b>3</b>
<b>Contents</b> .....	<b>4</b>
<b>1 Introduction</b> .....	<b>5</b>
1.1 WP7 – Veracity Intelligence for Patient Care.....	5
<b>2 Demonstration study 1 – Social Media and Medication Choices</b> .....	<b>6</b>
2.1 Aims.....	6
2.2 Results.....	6
2.2.1 Sentiment .....	6
2.2.1.1 Opinion Mining.....	6
2.2.1.2 Semeval-best .....	9
<b>3 Demonstration study 2 – Social media and ‘legal highs’</b> .....	<b>11</b>
3.1 Aims.....	11
3.2 Results.....	11
<b>4 Demonstration study 3 – Mental health stigma</b> .....	<b>12</b>
4.1 Aims.....	12
4.2 Methodology .....	12
4.3 Results.....	12
<b>5 Demonstration Study 4 – Self-harm and Suicide</b> .....	<b>18</b>
5.1 Aims.....	18
5.3 Annotation Process .....	18
5.3.1 Methodology .....	18
5.3.2 Results.....	19
<b>6 Dashboard Evaluation</b> .....	<b>20</b>
6.1 Methodology .....	21
6.1.1 Participants.....	21
6.1.2 Evaluation Tasks .....	24
6.1.3 Evaluation Sessions .....	25
6.2 Results.....	26
6.2.1 Task Completion Success Rate .....	26
6.2.2 Time to Completion .....	29
6.2.3 Overall Ratings .....	30
6.3 Recommendations and Conclusion.....	33
<b>References</b> .....	<b>35</b>
<b>Appendices</b> .....	<b>37</b>
Appendix 1 – Evaluation Tasks .....	37
Appendix 2 - Welcome and Purpose .....	39
Appendix 3 - User Impressions.....	40
Appendix 4 - System Usability Scale (SUS) .....	41

### **1 Introduction**

#### **1.1 WP7 – Veracity Intelligence for Patient Care**

The aim of WP7 was to utilize PHEME technologies ([www.pHEME.eu](http://www.pHEME.eu)) in practical applications for the mental healthcare domain. This will enable clinicians and public health professionals to explore online news and social media content for emerging mental-health related trends with a particular focus on rumours and misinformation. This evaluation will in turn be used (i) to develop educational materials for patients and the public, by addressing concerns and misconceptions, and (ii) to combine with analysis of data from the electronic health record in order to ascertain the potential impact of misinformation on clinical outcomes (e.g. whether periods of high stigmatizing content on social media coincide with higher rates of crisis episodes in those living with mental disorders). This case study has enabled the integration of project technologies into a clinical record application for the fundamental goal of monitoring mental-health related rumours and misinformation in online and social media.

In the first section of this deliverable, we summarize the progress made in each of our four case studies on psychotropic medication, legal highs, mental health stigma and self-harm and suicide.

In the second section, and in line with the work outlined in T7.4 ‘User-based Evaluation’, we summarise the deployment and evaluation of the interactive dashboard.

### **2 Demonstration study 1 – Social Media and Medication Choices**

#### **2.1 Aims**

The main aims of this study were to identify social media preferences, dislikes and rumours about certain medications in mental healthcare and how these relate to outcomes derived from the clinical record.

#### **2.2 Results**

As described in D7.2.2 ‘Annotated Corpus – Final Version’, we successfully developed a GATE application for identifying true instances of an advertisement in our medication tweet corpus with a precision score (positive predictive value) of 0.90 and a recall score (sensitivity) of 0.92. This application has been incorporated in the PHEME pipeline (D6.1.3 ‘PHEME Integrated Veracity Framework- v2.0) and implemented as an active filter in the dashboard’s drill-down menu as described in D5.3 ‘Usability Evaluation Report’.

##### *2.2.1 Sentiment*

We manually double-annotated 455 tweets relating to 5 medications for polarity (positive, negative, neutral) and subjectivity (subjective, objective) and ran the same tweets through three off-the-shelf algorithms: TextBlob (Loria et al, 2016), AFINN (Nielsen, 2011) and LabMT (Dodds et al, 2011). A preliminary investigation showed that there was very low agreement between the human annotators and the three algorithms. The main issue noted was the difficulty in the algorithms to distinguish between the tweet stance and the specific sentiment expressed toward the medication. Therefore, we explored two further algorithms.

##### *2.2.1.1 Opinion Mining*

With our colleagues in Sheffield University, we used an opinion mining application. For the experiment, we employed our standard generic rule-based opinion mining tool in GATE (Maynard, 2016) but with some modifications. The generic tool takes as input a set of possible candidate opinion targets, which are typically either terms or named entities. We therefore collected all the drug names and their variants and converted them to a case-insensitive gazetteer list, which was used to match against

### D7.3 Application and Evaluation Results

these terms found in the tweet text. Any term matched against this list was annotated as a potential candidate opinion target and fed into the opinion mining application. Second, the rules were modified slightly to ensure that only tweets that expressed sentiment about a particular drug (as matched against the list) were identified as positive or negative; the rest were identified as neutral even if they expressed sentiment.

We compared the system annotation with the gold standard provided on 6,885 annotations from the medication corpus, to check the accuracy. On the first pass, the tool achieved 56% precision on just the “relevant tweets” set, which we extracted from the larger set of annotated tweets. Note that a comparison against the entire set including the adverts would have generated higher accuracy since the proportion of neutral polarity would be higher overall (i.e. annotating neutral documents generally gets very high accuracy with our system).

Two issues influence this result:

1. Many of the gold standard annotations were found to be wrong, or unexpected, for example:

*RT @CraigyFerg: Tiny Wings is the methadone for my Angry Birds habit.*

➔ This is probably not a relevant tweet.

*Has anyone seen the girl in the jeans, on the Abilify commercial? Whoa!! Lol*

*abilify should give me the power to fly. false. advertising.*

➔ Both of these were annotated as negative, but should be neutral (the sentiment is about the advertisement not the drug).

*Diazepam is a beautiful thing.*

➔ This was annotated as neutral, should be positive.

2. We noted that there were also many tweets which were hard to interpret, even by a person. For these, an automated tool is also very likely to struggle.



### D7.3 Application and Evaluation Results

We also carried out a small experiment to evaluate the opinion target detection separately. The task was to detect whether people were talking positively or negatively about a particular drug. In this case, the name of the drug is already known, so it was not a task of target discovery, but only target assignment. The opinion target is the object of the opinion, e.g. in the phrase “I’m scared about Abilify”, and the author is showing negative sentiment (fear) about Abilify, so we would identify “Abilify” as the target of the opinion.

Unlike in the general sentiment task, here the system should not return neutral if a positive or negative sentiment is expressed about something other than the medication. For example, in the tweet:

*Avoid grapefruit juice and grapefruits when taking Latuda*

no opinion is expressed about Latuda, although a negative opinion is expressed about grapefruit and grapefruit juice (in the context of taking Latuda), so a correct response would be a neutral opinion. In this case, even though our system identifies both *grapefruits* and *Latuda* as possible target candidates, it does not assign the negative sentiment to the tweet. One drawback to the tool is that it does not perform nominal or pronominal co-reference between sentences, so if a target is mentioned explicitly in a previous sentence, it is not connected with the sentiment.

An example of this is the tweet:

*The most effective lucid dreaming supplement known to modern science is galantamine. This extract has become popular as a dream enhancer...*

Here, the first sentence is correctly analysed as expressing a positive opinion about galantamine. However, in the second sentence, while a positive opinion is correctly identified, no matching target is found.

In a sample subset of 50 random tweets from the “relevant tweets” set, 44 targets were correctly matched with the opinion (according to a manual inspection of the results). One target was missing (the case of the above tweet about galantamine), while 6 were identified incorrectly. For example in the tweet:

*Abilify commercials irritate me*

### D7.3 Application and Evaluation Results

“Abilify” was found as the target of the sentiment, whereas the correct target should have been “Abilify commercials” (and thus the sentiment was actually not relevant to the task and should not have been annotated). This problem could have been resolved with some improvements to the rules for candidate target detection.

Errors mostly occurred with our optional additional rules, where the target of the opinion is identified as a noun phrase, which has not been identified specifically as a candidate target (term or entity). For example, in the sentence “Ability bothers me”, “me” was wrongly identified as the target instead of “Ability”. These rules were added in order to try to improve recall, because we found many instances where the candidate target was not otherwise identified. Some improvements were therefore made to these rules in order to deal with these issues, and accuracy of the overall sentiment (as described above) increased to 61% as a result.

#### 2.2.1.2 Semeval-best

In an attempt to improve the precision of our sentiment detecting application, we decided, with our colleagues at Warwick University, to test another sentiment classification model on the same medication corpus of 7,000 annotated medication-related tweets (Wang et al, 2017). Semeval-best is a tweet-level sentiment classification model that uses extensive data preprocessing and feature engineering. Text pre-processing techniques include removing retweet “RT” and hashtag “#” symbols, removing URL links, normalising emoticons and abbreviations. It extracts various types of features from the tweets: (i) n-grams, (ii) lexicon features, (iii) word cluster features and (iv) word embedding features.

For n-grams, we used 1-2-3 grams after filtering out all the stop words then we converted the resulting feature matrix to tf-idf representations. We constructed 32 lexicon features from 9 Twitter specific and general-purpose lexica. Each lexicon provides either a numeric sentiment score or categories where a category could correspond to a particular emotion or a weak/strong positive/negative sentiment.

Word cluster features were derived from Brown Clusters (Gimpel et al, 2011). The use of word embedding features to represent the context of words and concepts has been shown to be very effective in boosting the performance of sentiment classification. Here, we used three different pre-trained word embedding resources

### D7.3 Application and Evaluation Results

including a sentiment sensitive embedding (Tang et al, 2014) for extracting features. Pooling functions *sum* and *average* are applied.

For classification, we used LIBLINEAR (Fan et al, 2008), which approximates a linear SVM. In optimising the cost factor  $C$  and the class weight parameter, we performed five-fold cross validation on the training data (70% of the whole data set) and selected the parameters that gave the highest accuracy score.

We followed previous work on Twitter sentiment classification and report our performance in accuracy, 3-class macro-averaged F1 score, 2-class macro-averaged F1 score, precision score and recall score (Table 1).

**Table 1** Performance metrics for Semeval-best

Accuracy	Accuracy	3-class f1	2-class f1	Precision	Recall
<b>Semeval-best</b>	87.56	52.00	31.38	77.26	46.74
<b>Target-ind</b>	85.96	35.83	7.54	81.05	35.94
<b>Target-dep+</b>	86.59	48.52	26.43	71.11	44.14
<b>TDparse</b>	87.03	48.32	25.97	74.54	43.93

Despite the higher precision achieved through this algorithm, the recall remained relatively low, most of the tweets were classified as neutral and it took over a minute for the model to run over each tweet.

The task of sentiment detection in medication-related tweets is hard for a variety of reasons. First, as mentioned above, it is often not clear even to a person what the target term is, or there may be multiple targets that could equally apply. Second, the span of the target has to be identified correctly and lastly, the target is sometimes implicit. As such, we decided not to further pursue the implementation of a medication-specific sentiment-detecting application in the dashboard and we maintained instead the use of the generic sentiment filter already developed by WP5 ‘Interactive Visual Analytics Dashboard’.

### **3 Demonstration study 2 – Social media and ‘legal highs’**

#### **3.1 Aims**

The main objectives of this study were to monitor the emergence of novel psychoactive substances in social media and the controversies surrounding them and to identify if and how promptly they appear in the mental health clinical record.

#### **3.2 Results**

We successfully developed a GATE application for automatically identifying genuine mentions of mephedrone in tweets with a precision score of 0.99 and a recall score of 0.90. The application<sup>1</sup> and the corpus<sup>2</sup> are available online.

Our manuscript reporting on the comparison of mephedrone mentions on Twitter, Wikipedia, Google and a large electronic mental health record database (CRIS) has been published in the journal *European Psychiatry* (Kolliakou et al, 2016).

Due to the paucity of references to legal highs in news and social media as well as in CRIS, the application has not been included in the PHEME pipeline or dashboard.

---

<sup>1</sup> [https://figshare.com/articles/Mephedrone\\_annotations\\_for\\_Twitter/1613832](https://figshare.com/articles/Mephedrone_annotations_for_Twitter/1613832)

<sup>2</sup> [https://figshare.com/articles/Mephedrone\\_annotations\\_for\\_Twitter/1613832](https://figshare.com/articles/Mephedrone_annotations_for_Twitter/1613832)

### **4 Demonstration study 3 – Mental health stigma**

#### **4.1 Aims**

The aim of this study was to identify how mental health stigma, as a particularly important instance of misinformation, arises in social media, the role rumours (as identified through stigmatising expressions) play in propagating these attitudes and how it relates to markers of negative impact measured in the clinical record.

#### **4.2 Methodology**

In addition to the development of an application for detecting anti-stigma through our Germanwings sub-study (D7.2.2), we also conducted an experiment on rumours conversations based on the same corpus.

Specifically, tweets associated with the Germanwings plane crash in March 2015 were collected by using Twitter's streaming API. We sampled the tweets associated with Andreas Lubitz, the co-pilot of the flight deemed responsible for the crash, who was rumoured to have been diagnosed with depression. We used the search terms \*depress\*, mental\* and psych\* to identify tweets related to the story. For each of these tweets, we collected all the replies following the methodology described in Zubiaga et al. (2016), forming conversations with the structure of a tree. The resulting 1,866 conversations were annotated as being initiated by a rumour or a non-rumour, because of our interest in focusing on rumours, and 1,596 of these conversations (85.5%) were manually identified as rumours. Out of those rumourous conversations, we randomly sampled 31 conversations for fine-grained annotation of each tweet in the conversation. Each of these tweets was annotated for support, evidentiality and certainty, as described in Zubiaga et al. (2016).

#### **4.3 Results**

The resulting collection is composed of 509 tweets that are part of these 31 conversations. These 509 tweets are distributed as follows: 31 are source tweets initiating the rumours, 237 tweets replying directly to source tweets, and 241 nested replies who respond to earlier replies. Table 2 shows the distribution of support, evidentiality and certainty as well as the number of direct and nested (deep) replies for each source tweet.

## D7.3 Application and Evaluation Results

**Table 2** Source tweets - support, certainty, evidentiality and type of reply

Tweet	Rumour	Support	Certainty	Evidentiality	Direct replies	Deep replies
1	Co-pilot suffered from depression	Supporting	Certain	None	13	8
2	Co-pilot's mental illness was responsible for the crash	Unclear	Somewhat-certain	Reasoning	5	15
3	Co-pilot had already undergone psychological testing	Supporting	Certain	Quoting-verifiable-source-url-given	2	9
4	Co-pilot's mental illness was responsible for the crash	Unclear	Certain	Reasoning	4	3
5	Psychiatric drugs to blame for crash	Unclear	Uncertain	Quoting-verifiable-source-url-given	4	18
6	Co-pilot's mental illness was responsible for the crash	Unclear	Somewhat-certain	None	3	9
7	Co-pilot's mental illness was responsible for the crash	Supporting	Certain	Quoting-verifiable-source-url-given	3	7
8	Co-pilot suffered from depression	Unclear	Underspecified	Reasoning	17	2
9	Co-pilot suffered from depression	Unclear	Certain	None	1	15
10	Co-pilot suffered from depression	Supporting	Certain	Quoting-verifiable-source-url-given	9	2
11	Co-pilot suffered from depression	Supporting	Somewhat-certain	Quoting-verifiable-source-url-given	13	1
12	Co-pilot suffered from depression	Supporting	Somewhat-certain	Quoting-verifiable-source-url-given	2	12
13	Co-pilot's mental illness was responsible for the crash	Denying	Certain	Reasoning	8	4
14	Co-pilot's mental illness was responsible for the crash	Denying	Somewhat-certain	Quoting-verifiable-source-url-given	1	11

### D7.3 Application and Evaluation Results

<b>Tweet</b>	<b>Rumour</b>	<b>Support</b>	<b>Certainty</b>	<b>Evidentiality</b>	<b>Direct replies</b>	<b>Deep replies</b>
15	Co-pilot's mental illness was responsible for the crash	Denying	Underspecified	Quoting-verifiable-source-url-given	5	12
16	Co-pilot suffered from depression	Supporting	Certain	Quoting-verifiable-source-url-given	12	0
17	Co-pilot's mental illness was responsible for the crash	Denying	Certain	Quoting-verifiable-source-url-given	1	18
18	Co-pilot suffered from depression	Unclear	Certain	None	3	12
19	Co-pilot hid mental illness from employers	Supporting	Certain	Quoting-verifiable-source-url-given	14	0
20	Co-pilot's mental illness was responsible for the crash	Denying	Certain	None	6	12
21	Co-pilot's mental illness was responsible for the crash	Unclear	Certain	Witnessed	10	4
22	Co-pilot's mental illness was responsible for the crash	Supporting	Somewhat-certain	Quoting-verifiable-source-url-given	9	8
23	Psychiatric drugs to blame for crash	Unclear	Underspecified	Quoting-verifiable-source-url-given	10	6
24	Co-pilot's mental illness was responsible for the crash	Unclear	Certain	Reasoning	1	8
25	Co-pilot's mental illness was responsible for the crash	Unclear	Certain	None	16	1
26	Co-pilot suffered from depression	Supporting	Certain	Quoting-verifiable-source-url-given	1	20
27	Co-pilot's mental illness was responsible for the crash	Unclear	Somewhat-certain	Quoting-verifiable-source-url-given	8	11
28	Co-pilot's mental illness was responsible for the crash	Unclear	Underspecified	Quoting-verifiable-source-url-given	12	1

### D7.3 Application and Evaluation Results

<b>Tweet</b>	<b>Rumour</b>	<b>Support</b>	<b>Certainty</b>	<b>Evidentiality</b>	<b>Direct replies</b>	<b>Deep replies</b>
29	Co-pilot's mental illness was responsible for the crash	Supporting	Certain	Quoting-verifiable-source-url-given	17	5
30	Psychiatric drugs to blame for crash	Supporting	Certain	Quoting-verifiable-source-url-given	13	7
31	Co-pilot's mental illness was responsible for the crash	Unclear	Underspecified	Quoting-verifiable-source-url-given	14	0



### D7.3 Application and Evaluation Results

Through our Germanwings case study, we also developed a GATE application for detecting anti-stigmating tweets with a 98% precision and 31% recall. This algorithm is part of the PHEME pipeline and an active filter in the dashboard's drill-down menu.

Our work on comparing mental health events in social media and CRIS is ongoing. Preliminary results show that a rise in anti-stigmatising tweets corresponds to an increase in inpatient and Home Treatment Team admissions in the clinical record regardless of year, season and bed occupancy. We will be shortly submitting our findings for peer-review publication.

Finally, we developed a focused version of the dashboard showing the number of documents relating to mental health issues as covered by online news media over a 2-month period for easy access by healthcare professionals (Figure 1). This is hosted on the *NIHR Maudsley Biomedical Research Centre* website.<sup>3</sup>

---

<sup>3</sup> [www.maudsleybrc.nihr.ac.uk/research/engagement-population-and-informatics/pHEME](http://www.maudsleybrc.nihr.ac.uk/research/engagement-population-and-informatics/pHEME)

## D7.3 Application and Evaluation Results

Maudsley  
Biomedical Research Centre  
Dementia Biomedical Research Unit

Search...

NHS  
National Institute for Health Research

HOME | ABOUT US | RESEARCH | PATIENTS & PUBLIC | TRAINING | PARTNERSHIPS | DEMENTIA BIOMEDICAL RESEARCH UNIT


BRC Home | Research | Engagement, population and informatics | PHEME

**PHEME – Computing Veracity Across Media, Languages and Social Networks**

This webpage hosts a series of live graphics showing the number of stories appearing in online news media sources, over the last two months, relevant to common mental health concerns. It aims to provide mental health professionals, service users and the general public with accurate information about recent online news media coverage.


The page has been developed as part of PHEME – a major international research project to understand how facts, half-truths and deception spread on the internet and how to counter this with better quality information. We need to identify stigma and misinformation in online news media in order to counter it and this webpage is part of that goal.

- What is PHEME?
- What is the role of the NIHR Maudsley Biomedical Research Centre?
- How do I use this site?
- Where can I get more information?




**Topic coverage**

READ MORE




**Dementia**

READ MORE



**Self Harm**

READ MORE



**Autism Spectrum Disorders**

READ MORE

Maudsley  
Biomedical Research Centre  
Dementia Biomedical Research Unit

Search...

NHS  
National Institute for Health Research

HOME | ABOUT US | RESEARCH | PATIENTS & PUBLIC | TRAINING | PARTNERSHIPS | DEMENTIA BIOMEDICAL RESEARCH UNIT

BRC Home | Research | Engagement, population and informatics | PHEME | PHEME - Topic Coverage

Leadership Group A-Z
NIHR Senior Investigators
Clinical disorders
Engagement, population and informatics
Bioinformatics and statistics
Clinical and population informatics
Patient and carer participation
PHEME
PHEME - Dementia
PHEME - Bipolar Affective Disorder
PHEME - Self Harm
PHEME - Autism Spectrum Disorders
PHEME - Schizophrenia/Psychotic Disorder
PHEME - Attention Deficit Disorder
PHEME - Depressive Disorder
PHEME - Alzheimer's Disease
PHEME - Obsessive-compulsive Disorder
PHEME - Anxiety Disorder
PHEME - Topic Coverage

**Topic Coverage**

This representation gives an overview of the media exposure that each of the 11 mental health concerns cover – it represents the number of stories relevant to each of our topics as a proportion of the total number of stories relevant to the 11 topics and published in the two months indicated in the individual topic graphs. Please place the cursor on the chart to see the number of news stories related to each mental health concern during a particular week, and the top 3 terms associated with these. Clicking on any section of the chart will give you an option to 'Explore Documents'. Clicking on this option will take you to the main PHEME dashboard where you can access the news stories corresponding to the mental health concern of the section clicked.

*topic title*

**Dementia**

Bipolar Affective Disorder

Self Harm

Autism Spectrum Disorders

Schizophrenia/Psychotic Disorder

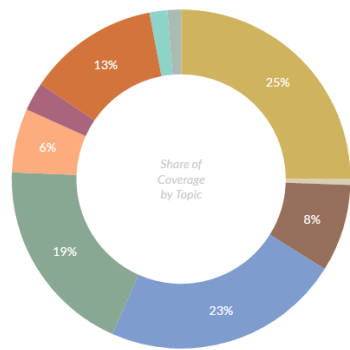
Attention Deficit Disorder

Depressive Disorder

Alzheimer's Disease

Obsessive-compulsive Disorder

Anxiety Disorder



Topic	Share of Coverage
Dementia	25%
Autism Spectrum Disorders	19%
Depressive Disorder	23%
Alzheimer's Disease	8%
Attention Deficit Disorder	6%
Obsessive-compulsive Disorder	13%
Anxiety Disorder	8%

Figure 1 Screenshot of the PHEME news media monitoring tool and the topic coverage option

### **5 Demonstration Study 4 – Self-harm and Suicide**

#### **5.1 Aims**

The main objective of this study was to describe online ‘chatter’ around themes of self-harm and suicide including exploration of rumours surrounding suicide deaths of famous individuals and potential relationships with the clinical presentations of vulnerable patient groups.

#### **5.3 Annotation Process**

##### *5.3.1 Methodology*

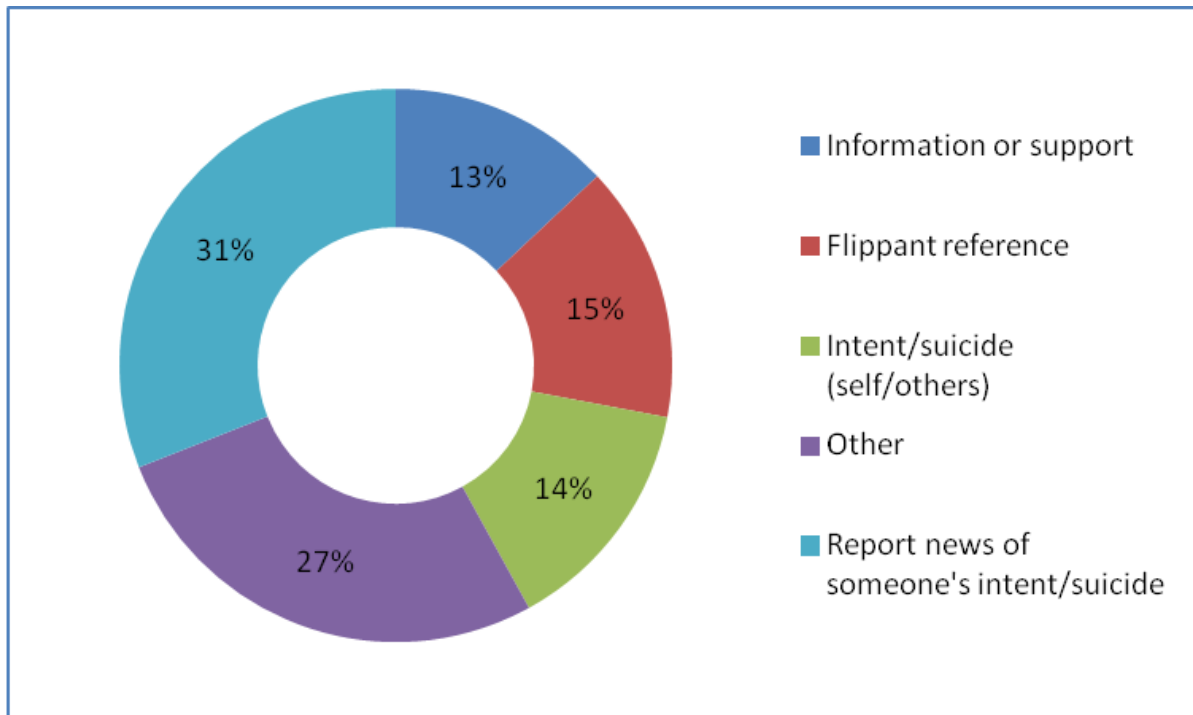
As described in D7.2.2, we performed a two-part annotation on the Twitter corpus coding tweets based on relevancy, subjectivity and whether they expressed a personal experience or opinion. After preliminary analysis, it became apparent that the nature of self-harm and suicide-related tweets made it difficult to distinguish them on the basis of subjectivity and experience or opinion. In order to improve and make the classification more meaningful, two annotators manually re-annotated a proportion of the extracted tweets based on relevancy and then further annotated a proportion of the relevant tweets into the following categories: news of someone’s intent/suicide, intent/suicide (self/other), flippant reference, information or support, or other. This annotation schema was based on the one developed by Colombo et al (2016). An inter-annotator agreement analysis using the Kappa statistic was performed to determine consistency between the two annotators.

To develop an automatic application for identifying relevant references to self-harm and suicide in the remaining tweets, a natural language processing (NLP) approach was taken. This involved applying an algorithm that was able to determine if the semantic meaning of the text was relevant to self-harm and suicide (Wang et al, 2017). The algorithm was developed using a subset of the tweets already annotated. These tweets were analysed and the linguistic patterns that indicated a relevant reference to self-harm and suicide were determined, which were then used to create identification rules implemented using the General Architecture for Text Engineering software (GATE; Cunningham et al, 2011). GATE also supported the rapid deployment of these applications over the larger set of tweets retrieved. Rules were tested over another ‘gold standard’ subset of tweets already annotated.

## D7.3 Application and Evaluation Results

### 5.3.2 Results

A first annotator coded 1,094 tweets, of which 300 were double-annotated ( $Kappa=0.90$ ). Figure 2 shows the proportion of relevant tweets (673) corresponding to each annotation code.



**Figure 2** Proportion of self-harm and suicide-related tweets corresponding to the five annotation codes

An NLP algorithm was developed through the use of 360 annotated tweets (training set). The rules created to identify the linguistic patterns indicating a relevant reference to self-harm and suicide were then tested on another 360 annotated tweets (gold standard set) using GATE. The development of the GATE application was successful in identifying relevant references to self-harm and suicide in the tweets with a precision score of 0.92 and a recall score of 0.58. The application was then deployed over the complete dataset of 78,149 tweets retrieved between 2009 and 2014 – 63,097 were identified as relevant references to self-harm and suicide. This application has been included in the PHEME pipeline and serves as a filter in the dashboard's drill-down menu.

Our findings from the Reddit analysis of rumorous conversations regarding highly-publicised deaths together with the results from the experiment on the Germanwings-related rumours in Twitter (D7.2.2) are currently being prepared for publication submission.

### 6 Dashboard Evaluation

The aim of WP7 is to tailor and adopt PHEME technologies in practical application for the mental healthcare domain. This will enable clinicians and public health professionals to explore online news and social media content for emerging mental-health related trends with a particular focus on rumours and misinformation. This evaluation will in turn be used (i) to develop educational materials for patients and the public, by addressing concerns and misconceptions, and (ii) to combine with analysis of data from the electronic health record in order to ascertain the potential impact of misinformation on clinical outcomes (e.g. whether periods of high stigmatizing content on social media coincide with higher rates of crisis episodes in those living with mental disorders). This case study has enabled the integration of project technologies into an application for the fundamental goal of monitoring mental-health related rumours and misinformation in online and social media.

The visual analytics tool developed by WP5, hereafter referred to as the ‘dashboard’, have provided the interactive interface through which mental-health related rumours and information across news media and social networks can be explored. In line with the work outlined in T7.4, this section summarises the deployment and evaluation of the medical dashboard.

The usability tests were conducted with a view to evaluate users’ experience of individual components developed in WP7 as well as satisfaction with general interactive exploration capabilities and functions. To this end, three types of usability assessments were conducted:

- **Formative usability testing** – on-site, 1-2-1 sessions of audio-recorded observations of users working on a set of real-life pre-defined tasks.
- **Qualitative evaluation** – post-session, open-ended questions on user impressions of general interface usage.
- **Quantitative evaluation** – post-session completion of standardised, 10-item System Usability Scale (SUS).

## D7.3 Application and Evaluation Results

### **6.1 Methodology**

#### *6.1.1 Participants*

All attendees had a professional connection to the NIHR Maudsley Biomedical Research Centre. Six participants in total attended on three consecutive testing dates: two on Thursday 18<sup>th</sup> January, three on Friday 19<sup>th</sup> January and one on Saturday 20<sup>th</sup> January 2017. Demographics and background information are presented in Table 3. Participants were mostly male and white. We ensured representation from across age groups and different professions. Users reported using the computer for a variety of reasons and only one attendee indicated having advanced technical knowledge.

D7.3 Application and Evaluation Results

**Table 3** Participant demographics and background information

		User 1	User 2	User 3	User 4	User 5	User 6
<b>Age</b>	16-24						✓
	25-34			✓			
	35-44		✓		✓		
	45-54					✓	
	55-64						
	65+	✓					
<b>Gender</b>	Female			✓			✓
	Male	✓	✓		✓	✓	
<b>Ethnicity</b>	Asian						✓
	White	✓	✓	✓		✓	
	Black						
	Mixed						
	Other						
<b>Profession</b>	Academic / Clinician		✓	✓	✓		
	Service User/Carer Representative in Research	✓				✓	
	High School Student						✓

D7.3 Application and Evaluation Results

**Table 3** Participant demographics and background information cont.

		User 1	User 2	User 3	User 4	User 5	User 6
<b>Reasons for computer use</b>	Gaming/entertainment						
	Reading the news		✓	✓	✓	✓	
	Shopping/banking	✓	✓	✓	✓	✓	
	Programming/coding				✓		
	Microsoft Office activities		✓	✓	✓	✓	✓
	Social Networking			✓	✓		✓
	Other	✓				✓	
<b>Daily hours spent on computer</b>	0-2						
	3-5	✓					
	4-6						✓
	7-10			✓	✓		
	More than 11		✓			✓	
<b>Advanced technical knowledge</b>	Yes				✓		
	No	✓	✓			✓	✓
	Not strictly, but more than average			✓			



### 6.1.2 Evaluation Tasks

The evaluation tasks were created by the WP7 team to comprise a number of different scenarios, features and functionalities. Our aim was to develop the tasks in a manner that replicated investigating trends and rumours by navigating a large part of the dashboard, as would happen in real life. On completion of the tasks, the participants were to have had the opportunity to practice identifying and exploring trends and rumours by:

- 1) Choosing media sources and focusing the search on a particular mental health issue
- 2) Utilising the visualizations to discover how the topic of interest relates to other stories, locations and sentiment
- 3) Retrieving the stories relevant to the search topic and its associations
- 4) Comparing the media coverage between the topic of interest and other mental health issues
- 5) Determining the type and quality of the sources that publish the stories
- 6) Isolating peaks in coverage and locating the news stories relating to them
- 7) Combining searches with two or more topics of interest to uncover precise content
- 8) Using anti-stigma and veracity classifications as a proxy for biased and rumorously stories

Subsequently, a total of 11 tasks were developed to test use of sources and configuration, trend charts, content view (search results), associated terms, visualizations (map, tag cloud and keyword graph), temporal controls and synchronization and tooltips. In addition, we devised 8 further tasks to test the components of particular interest to WP7: sentiment, anti-stigma and veracity, available through the drill-down menu.

Similar to the internal evaluations performed by WP5, each task was rated by the test administrator as 0, 1 or 2 corresponding to non-completion, completion with difficulty or help, and easy completion. Notes were also taken by the administrator based on actions performed and thoughts expressed by the participant (e.g. wrong pathway, confusing page layout, navigation issues and terminology) as seen in example of task 1 in Figure 3. These

notes were cross-referenced with the audio-recordings to develop an evaluation narrative reported in the subsequent sections. The complete set of tasks can be viewed in Appendix 1.

**Task 1– Choose source**

*From the sources, choose ‘News Media’, ‘Google’ and ‘Twitter’.*

Pathway(s)	Success	Notes/Observations
	0 Not completed	Wrong pathway Confusing page layout Navigation issues Terminology
	1 Completed with difficulty or help	
	2 Easily completed	

**Figure 3** Example of task evaluation scoring sheet for use by task administrator

*6.1.3 Evaluation Sessions*

The WP7 team contacted and recruited participants through professional networks. Emails were sent to attendees informing them of evaluation session logistics and requesting their availability and participation.

Each individual session lasted up to one hour. At the beginning of the session, the test administrator read out a standard ‘Welcome and Purpose’ information sheet (Appendix 2) and gave participants the opportunity to ask questions. Additionally, they were invited to complete a brief demographics and background questionnaire.

Then, attendees were asked to participate in a training session in the form of the screencast created by WP5 and described in D5.3. Following the video, participants were provided with a two-sided A4 overview of the main dashboard sections and features for reference, taken from the ‘Help’ section<sup>4</sup> of the dashboard, if and when needed, and evaluation commenced. The test administrator read out each of the tasks to participants, encouraged them to work

<sup>4</sup> [www.weblyzard.com/interface/](http://www.weblyzard.com/interface/)

through the tasks on their own as much as possible and to talk aloud in describing how they were navigating the dashboard and the actions they were performing (Appendix 1). If after considerable effort the participant was unable to complete the task described, the test administrator provided tips or direction for successful task completion. Time to completion of all tasks was noted and all evaluations were audio-recorded.

At the end of the evaluation session, participants were invited to share their thoughts and opinions on the experience of using the dashboard based on 7 open-ended questions (Appendix 3) and a 10-item SUS (Sauro, 2011; Appendix 4). SUS yields a single number representing a composite measure of the overall usability of the dashboard. To calculate the SUS score, we first summed the score contributions from each item. Each item's score contribution ranged from 0 to 4. For items 1, 3, 5, 7 and 9 the score contribution is the scale position minus 1. For items 2, 4, 6, 8 and 10, the contribution is 5 minus the scale position. We then multiplied the sum of the scores by 2.5 to obtain the overall value of system usability. SUS scores range from 0 to 100.

Two trial sessions were also held with the assistance of colleagues from the NIHR Maudsley BRC to test and fine-tune the tasks and session procedures prior to the official evaluations.

## **6.2 Results**

### *6.2.1 Task Completion Success Rate*

The test administrator recorded participants' ability to complete the tasks and kept note of comments, requests for assistance and wrong navigational or pathway choices. The administrator provided tips or guidance only after considerable effort by the participant was unsuccessful or there was a request for assistance. As such, none of the tasks were left uncompleted.

Tasks 1, 4, 9, 13, 15, 17 and 18 were completed with ease by all participants. Tasks 2 and 7, tasks 3, 5, 12, 14, 16 and 19 and tasks 6 and 8 were easily completed by 83%, 67% and 50% of participants, respectively. Finally, only 33% of participants easily completed tasks 10 and 11. Individual task completion rates are shown in Table 4. The majority of users were able to easily complete over 2/3 of the tasks.

**Table 4** Task completion rates

Task	User 1	User 2	User 3	User 4	User 5	User 6	% of users completing with ease
1	2	2	2	2	2	2	100%
2	1	2	2	2	2	2	83%
3	1	2	2	2	2	1	67%
4	2	2	2	2	2	2	100%
5	1	1	2	2	2	2	67%
6	2	2	1	2	1	2	50%
7	1	2	2	2	2	2	83%
8	1	2	2	1	2	1	50%
9	2	2	2	2	2	2	100%
10	1	1	2	1	1	2	33%
11	1	1	1	2	1	2	33%
12	1	2	2	2	1	2	67%
13	N/A	2	2	2	N/A	N/A	100%
14	N/A	1	2	2	N/A	N/A	67%
15	N/A	2	2	2	N/A	N/A	100%
16	N/A	2	1	2	N/A	N/A	67%
17	N/A	2	2	2	N/A	N/A	100%
18	N/A	2	2	2	N/A	N/A	100%
19	N/A	2	1	2	N/A	N/A	67%
<b>% of tasks completed with ease</b>	<b>33%</b>	<b>79%</b>	<b>79%</b>	<b>90%</b>	<b>67%</b>	<b>83%</b>	

The issues encountered/raised by participants during the tasks completed with difficulty are summarized in Table 5.

**Table 5** Task and common issues description

Task	Task description / issue
2	<p><b>Run search for ‘Dementia’ from the mental health disorders topics</b></p> <p>Issue: Confusion between activating a search and activating the topic in the trend chart.</p>
3	<p><b>Identify the first 3 associations with ‘Dementia’ and indicate at least one visualization where these associations are shown</b></p> <p>Issue: Uncertainty between hierarchy of associations in terms of order and number of documents relating to each term.</p>
5	<p><b>Show the sources of the search results and sort them by Reach (high to low). Which is the top source in terms of reach? Is the sentiment relating to the search term from the first 5 sources mostly negative or positive?</b></p> <p>Issue: Difficulty in activating side-menu by hovering over search results as well as difficulty in locating source list. No issues with identifying sentiment values.</p>
6	<p><b>Adjust the keyword tree to show the top 6 associations with the search term</b></p> <p>Issue: Most participants changed the edges to 6 but couldn’t remember how to refresh the graph so that new number of associations is shown.</p>
7	<p><b>Activate the first 5 mental health disorders (including dementia) in the trend chart – i.e. those that show the highest number of mentions. What percentage of coverage corresponds to each disorder?</b></p> <p>Issue: Difficulty in distinguishing between activating search and activating terms in trend chart</p>
8	<p><b>Identify the date range for the highest peak in ‘Schizophrenia/Psychotic Disorders’ in the trend chart and name the top 3 keywords associated with ‘Schizophrenia/Psychotic Disorders’ during that period.</b></p> <p>Issue: Small size and font in chart made it difficult to identify peak, date range and keywords.</p>
10	<p><b>Search for the terms ‘Dementia’ AND ‘Risk’ using the tooltip function. How many documents are retrieved for this search?</b></p> <p>Issue: Difficulty in locating tooltip. Confusion over different search options e.g. expand vs restrict</p>

**Table 5** Task and common issues description cont.

Task	Task description / issue
11	<p><b>Change the date interval to January 01 2017 – January 15 2017. How many documents in total are available during this time?</b></p> <p>Issue: Participants attempted to change date using only the right-hand side calendar. They needed prompting to choose the first date from calendar on the left. They also couldn't recall how to search for all available documents.</p>
12	<p><b>What percentage of all documents during this period shows positive, negative and neutral sentiment?</b></p> <p>Issue: Difficulty in activating side-menu by hovering over trend chart.</p>
14	<p><b>There is an increase in tweets related to ASDs between July 16<sup>th</sup> 2016 and July 26<sup>th</sup> 2016. Search for the terms 'Autism Spectrum Disorders' AND 'Therapist' using the tooltip function. Looking at the first 10-15 documents retrieved, can you very briefly describe the news story associated with this peak?</b></p> <p>Issue: Difficulty in locating tooltip.</p>
16	<p><b>Using the drill-down menu, identify whether the sentiment related to this search is mostly positive, negative or neutral.</b></p> <p>Issue: Difficulty in locating drill-down menu.</p>
19	<p><b>Run search for 'Anti-stigma'. How has the average sentiment changed?</b></p> <p>Issue: Confusion between activating search and activating term in trend chart.</p>

### 6.2.2 Time to Completion

The administrator recorded the time to completion for each participant. Individual completion times are shown in Table 6. Participants in the shorter evaluation sessions (in bold) completed the first 11 tasks in 16', on average. Those in the extensive sessions, averaged 22' for completion of 19 tasks.

**Table 6** Time to completion

User	Time
1	24'
2	26'
3	24'
4	16'
5	14'
6	10'

### 6.2.3 Overall Ratings

#### Quantitative Evaluation

After session completion, participants rated the dashboard on a 10-item SUS. The received feedback in the form of comments documents a generally positive impression but also shows that the complexity of the dashboard can be overwhelming for first-time users (individual numeric scores are presented in Table 7). Therefore, proper training is essential and impacts the perceived usability of the dashboard. While the opportunity to discuss certain tasks and ask questions during the session had a positive impact on the score, for example, as compared to previous internal evaluation where test users had little training and no opportunities to ask questions, the results also underscore that a 20-minute video introduction cannot fully replace a two-hour personal workshop offered to professional analysts (the format recommended by WP5 and one that the PHEME consortium will adopt as part of the follow-up exploitation activities planned in WP9).

**Table 7** System Usability Scores

External User	Score
1	55
2	62.5
3	55
4	62.5
5	55
6	77.5
<b>Average score</b>	<b>61.25</b>

## Qualitative Evaluation

Upon completion of the tasks, participants provided feedback on 7 questions addressing a range of subjects from most to least liked dashboard feature to recommendations for improvement. The responses to each question are summarized below.

### 1) What is your overall impression of the dashboard?

Positive	Negative
Could be easy to use with practice	Initially overwhelming
Hypothesis generating + explorative	Needs getting used to
Combination of different graphics that look at associations	Complicated/ technical terminology
Robust, extensive tool	A bit intimidating
Impressive in collecting all this information	Not intuitive
Different data sources	Have to move mouse around a lot to see different menus
So much information in one place	
Easy interface to use for something so complicated	

### 2) What did you like best about the site?

Information about a lot of topics
Worldwide coverage
Intuitively easy trends to read
All information on one page
Drill-down menu
So many options
Automated so no need to guess about associations and relevant content
Very good at telling you what is in the news/ positive-negative/what it is linked to
Visualizations to help quickly see what's important and whereabouts it is happening

### 3) What did you like least about the site?

Unfamiliar terminology
A lot on one page – maybe tabbing/personalising the page to increase font size
Keyword graph difficult to understand
Run search/activate in trend chart confusing, breaks consistency, not easy to understand



Make document list shorter
----------------------------

**4) If you were the website developer, what would be the first thing you would do to improve the website?**

Would make better use of colours – a little confusing
---

Layout/lack of tabs/small font size
-------------------------------------

Make drill-down menu more visible – it's the most important thing
---

Symbols to say there is a menu available when hovering over
---

Change the hover-over to normal drop down menu – they overlap and get in the way
--

**5) Is there anything that you feel is missing on this site?**

Population metrics
--------------------

Video instructions/ examples of functions
---

Summary of the top news/ what's trending
--

More sophisticated social network analysis behind the content generated
---

**6) If you were to describe this site to a friend or colleague in a sentence or two, what would you say?**

Dashboard which uses a number of programs to get current info from media sources / what words are being used in relation to mental health disorders
---

All publically available content on web that uses a tool to select defined searches, look at patterns and answers where things come from
--

Media tool but bigger and more powerful for monitoring news and media about mental health subjects and analyse them in detail
---

Tool that allows users to query and understand the buzz behind social/news media data sources based on temporal characteristics and other options
---

Program for analysis of news on mental health to show if coverage is positive/negative for different types of mental health issues and where it is being reported
---

Website that helps determine whether news articles are true or not. Useful for those needing information for research – make life easier
--

**7) Do you have any other final comments or questions (if different than above)?**

Wide perspective
------------------

Useful info for researchers
-----------------------------

Is there potential for data linkage to other databases?
---

Present it to media relations people/ science communications working in mental health
---

Visible description of the coverage of the media content/data sources
---

## **6.3 Recommendations and Conclusion**

We present here three core proposed changes driven by the participant success rate, behaviours and comments. Each recommendation summarizes the challenge observed with suggestions for improvement or resolution. We hope these changes will address areas where participants experienced problems or found the interface architecture unclear and enhance the overall ease of use (Figure 4).

### **1) Activating a search and activating a term in the trend chart**

Several participants had difficulties differentiating between the two. We suggest that training material clearly separates these two functions and presents scenarios where each is employed.

### **2) Activating side-menus**

Users found the side-menus in the trend chart and content view problematic. On one hand, they thought it was hard to remember how to activate them when needed and also, they considered them intrusive when trying to reach another function as they pop up regardless. We advise the menus be created in a traditional, stationary format.

### **3) Locating the tooltip and drill-down menu**

Participants considered these features to be two of the most important. However, they believed they were almost hidden and navigating to them was unclear. In the future, these two options should be placed in an obvious position on the interface.

Overall, the evaluation showed that users regarded the dashboard as a robust and extensive tool that collects large amounts of information from many different sources. In particular, the options available through the drill-down menu – stance, sentiment, veracity and anti-stigma – were highly commended by academics and clinicians. Participants highlighted the hypothesis-generating and research potential of such an interface. Those who experienced obstacles in accomplishing the tasks believed that ease of use could be achieved with training and practice. As with observations from the evaluation described in D5.3 and in line with the aim of the interface to serve as a tool for professionals, users with a technical background or more online practice performed better and favoured the dashboard as intuitive. Participants thought that the worldwide coverage, visualizations and ability to quickly identify important mental health topics and discussions were the major strengths of the interface.



**Figure 4** Screenshot of the dashboard with the three areas for improvement indicated

On average, users shared similar feedback on how to improve the layout and interactivity of the dashboard. There were suggestions for improving colour coding, increasing the font size, and revising some of the menu options in terms of clarity and visibility. In more detail, participants reported it would be helpful to include a summary of top news and trends, and to include a more sophisticated social network analysis. They would also welcome more detailed instructions for use, perhaps in the form of short videos covering different scenarios.

Evident in the feedback received, the dashboard represents a powerful tool for mental health media monitoring, which would benefit from the suggested modifications outlined in the previous section. Together with the main recommendations, secondary suggestions will be assessed for feasibility and implementation. We continue to explore the prospect of the tool in aiding the work of UK charities and are in close collaboration with *MIND* and *Time To Change* to support them in using the news media monitoring tool on our BRC website to achieve this potential. Further, we are planning a demonstration/introduction event for PR and communications teams working in science. Finally, we maintain strong links with the wider PHEME group and our representatives to continuously evolve the interface.

## References

- Colombo BG, Burnap P, Hodorog A, Scourfield J. Analysing the connectivity and communication of suicidal users on twitter. *Computer Communications* 2015; 73: 291-300.
- Cunningham H, Maynard D, Bontcheva K on behalf of the GATE group. *Text processing with GATE (Version 6)*. University of Sheffield, 2011.
- Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth MC. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. *PLoS One* 2011; 6(12): e26752.
- Fan R-E, Chang K-W, Hsieh JC, Wang X-R, Lin CJ. LIBLINEAR: a library for large linear classification. *J Machine Learning Res.* 2008; 9(9): 1971-4.
- Gimpel K, Schneider N, O'Connor B, Das D, Mills D, Eisenstein J, Heilman M, Yogatama D, Flanigan J, Smith NA. Part-of-speech tagging for Twitter: Annotation, features, and experiments. *Proceedings of the Conference of the Association for Computational Linguistics*, 2011.
- Kolliakou A, Ball M, Derczynski L, Chandran D, Gkotsis G, Deluca P, Jackson R, Shetty H, Stewart R. Novel psychoactive substances: An investigation of temporal trends in social media and electronic health records. *Eur Psychiatry* 2016; 38: 15-21.
- Loria S, Keen, P, Honnibal, M, Yankovsky R, Karesh D, Dempsey E, Childs W, Schnurr J, Qalieh A, Ragnarsson L, Coe J, Calvo AL. *TextBlob: Simplified Text Processing*. 2016. Available from <https://textblob.readthedocs.io/en/dev/>
- Maynard, D. Environmental Opinion Extraction. *DecarboNet Project Deliverable 2.3.2*, 2016.
- Nielsen FA. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs, *Proceedings of the Conference on European Semantic Web, 2011*.
- Sauro, J. (2011). *A Practical Guide to the System Usability Scale: Background, Benchmarks, and Best Practices*. Denver, USA: Measuring Usability LLC.

Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B. Learning sentiment-specific word embedding for twitter sentiment classification. *Proceedings of the Conference of the Association for Computational Linguistics*, 2014.

Wang B, Liakata M, Zubiaga A, Procter R. TDParse-multi-target-specific sentiment recognition on Twitter. *Proceedings of the Conference of the Association for Computational Linguistics*, 2017.

Zubiaga A, Liakata M, Procter, R, Hoi GWS, Tolmie P. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS One* 2016; 11(3): e0150989.

## Appendices

### Appendix 1 – Evaluation Tasks

1. From the sources, choose ‘News Media’, ‘Google’ and ‘Twitter’.
2. Run search for ‘Dementia’ from the mental health disorders topics.
3. Identify the first 3 associations with ‘Dementia’ and indicate at least one visualisation where these associations are shown.
4. Show the full text for the second document retrieved. Return to document list view.
5. Show the sources of the search results and sort them by Reach (high to low). Which is the top source in terms of reach? Is the sentiment relating to the search term from the first 5 sources mostly negative or positive?
6. Adjust the keyword tree to show the top 6 associations with the search term.
7. Activate the first 5 mental health disorders (including dementia) in the trend chart – i.e. those that show the highest number of mentions. What percentage of coverage corresponds to each disorder?
8. Identify the date range for the highest peak in ‘Schizophrenia/Psychotic Disorders’ in the trend chart and name the top 3 keywords associated with ‘Schizophrenia/Psychotic Disorders’ during that period.
9. Maximise the window of the geographic map. Which mental health disorder is associated with Hong Kong? Minimise the geographic map window.
10. Search for the terms ‘Dementia’ AND ‘Risk’ using the tooltip function. How many documents are retrieved for this search?
11. Change date interval to January 01 2017 – January 15 2017. How many documents in total are available during this time?
12. What percentage of all documents during this period shows positive, negative and neutral sentiment?  
[Choose July 7 to July 31 2016 as the date range. Choose Nice as the data source] – Administrator Task
13. Run search for ‘Autism Spectrum Disorders’ from the mental health disorders topics and activate the first 5 mental health disorders in the trend chart (incl. ASD) - – i.e. those that show the highest number of mentions.
14. There is an increase in tweets related to ASDs between July 16<sup>th</sup> 2016 and July 26<sup>th</sup> 2016. Search for the terms ‘Autism Spectrum Disorders’ AND ‘Therapist’ using the tooltip function. Looking at the first 10-15 documents retrieved, can you very briefly describe the news story associated with this peak?

15. Return search to 'Autism Spectrum Disorders' only.
16. Using the drill-down menu, identify whether the sentiment related to this search is mostly positive, negative or neutral.
17. Scroll down the drill-down menu to reveal all filters. How many documents have a verification status of "true"?
18. How many anti-stigmatising tweets have been retrieved?
19. Run search for 'Anti-stigma'. How has the average sentiment changed?

## **Appendix 2 - Welcome and Purpose**

Thank you so much for coming in today. I wanted to give you a little information about what you will be looking at and give you time to ask any questions you might have before we get started.

Today we are asking you to serve as an evaluator of an interactive dashboard and to complete a set of tasks. I am going to be asking you to look for some information on the dashboard and tell me how easy or difficult it was to find the information. These activities are all about how easy we have made it for people to use the dashboard.

During the session, I would like you to think aloud as you work to complete the tasks. I may ask you to clarify what you have said or ask you for information on what you were looking for or what you expect to have happen. I would like you to complete the tasks on your own as much as possible. If after a considerable effort you feel assistance is necessary, I will be able to answer questions or guide you towards a resolution. If you ever feel that you are lost or cannot complete a task with the information that you have been given, please let me know. I will then either put you on the right track or move you on to the next scenario.

I am here to record your reactions and comments of the dashboard you will view. We will be audio-recording this session for reference. We are only capturing your voice and your name will not be associated or reported with data or findings from this evaluation.

I may ask you other questions as we go and we will have wrap up questions at the end.

Is there anything you would like to ask before we begin?



## **Appendix 3 - User Impressions**

1. What is your overall impression of the dashboard?
2. What did you like best about the site?
3. What did you like least about the site?
4. If you were the website developer, what would be the first thing you would do to improve the website?
5. Is there anything that you feel is missing on this site?
6. If you were to describe this site to a friend or colleague in a sentence or two, what would you say?
7. Do you have any other final comments or questions?

## Appendix 4 - System Usability Scale (SUS)

**User name:**

Please state how much you agree with each statement on the left on a scale of 1-5 (strongly disagree to strongly agree)

		<b>Strongly disagree</b>	<b>Disagree</b>	<b>Neither agree nor disagree</b>	<b>Agree</b>	<b>Strongly agree</b>
1	I think that I would like to use this system frequently	1	2	3	4	5
2	I found the system unnecessarily complex	1	2	3	4	5
3	I thought the system was easy to use	1	2	3	4	5
4	I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
5	I found the various functions in this system were well integrated	1	2	3	4	5
6	I thought there was too much inconsistency in this system	1	2	3	4	5
7	I would imagine that most people would learn to use this system very quickly	1	2	3	4	5
8	I found the system very cumbersome to use	1	2	3	4	5
9	I felt very confident using the system	1	2	3	4	5
10	I needed to learn a lot of things before I could get going with this system	1	2	3	4	5