The search for credibility in news articles and tweets

An explorative journey to discover credibility features in whaling events

Marc Jacobs (2533914)
Information Sciences
Business Information Systems
VU University Amsterdam
marc@mvjacobs.nl

ABSTRACT

The introduction of social media as news source makes the credibility of online news questionable. This research aims at the discovery of credibility features in online news. Tweets and online news articles concerning whale hunting are gathered from the internet and the body and metadata were investigated for credibility features. Three crowdsourcing tasks were conducted where crowd workers (crowd users) compared two tweets or articles and indicated which one was more credible, and what features indicated this credibility. The results show that for tweets the author followers, and number of Wikipedia entities correlate strongly ($r > 0.6$) with the credibility of the tweet. Five crowd workers indicated that human-like stats and sentiment are important factors for the credibility assessment of a tweet. News articles do not have strong correlated features with credibility. Four crowd workers indicated that the completeness of an article is an important credibility factor.

Keywords

Credibility, tweets, articles, news, whaling, natural language processing, crowdsourcing, data mining

1. INTRODUCTION

The large amount of unstructured data on the Web causes an explosion of information, making it difficult for people to find what they are looking for. People are allowed to upload what they want, resulting in an impenetrable pile of information, consisting of duplicate and unreliable articles, blogs and other web media **[1]**. In 2008, Twitter had between 6 and 7 million active users **[13]**, which currently have been grown to 302 million active users, sending about 500 million tweets a day on the network **[14]**. A commercial news aggregator, Google News, has more than 4.000 news sources. Yahoo News aggregates from more than 5.000 content publishers. All of them acknowledge the need for content curation **[18]**.

For these vast amounts of unstructured data and online information, data science has become an important area of research. In the past decades discoveries like natural language processing **[2]** and event detection in news **[3]** allowed us to analyse the past, but more important, to predict the future based on the past.

One of the main accelerators behind these gigantic amounts of data is the notion of web democracy **[20]**. Everybody can post anything anytime. Naturally, this has a negative impact on the quality assurance of the posted items, since it is impossible to check all these items manually. The overwhelming amount of posts inquire a new type of quality assessment. Several studies **[11, 33, 34]** created credibility algorithms, using machine learning, that try to automatically determine the credibility of tweets. In these studies the focus on the features of the news sources was limited.

This research aims at the discovery of credibility features in tweets and online articles (professional news articles, opinion pages and blogs from The Guardian and The New York Times).

The context of this research is constrained to "whaling" or "whale hunting", because of the controversy of the topic, which introduces many facts and much fiction, referring to a balanced amount of credible and non-credible statements. The tweets and articles relevant to the whale hunting event are obtained and ranked based on the relevancy to the whaling event.

One part of the research concerns the development of a Credibility Framework. Credibility is defined in paragraph 3.1 as a combination of competence, trustworthiness and goodwill **[21]**. Using this definition of credibility the literature is searched for generic credibility factors, which eventually are used to derive concrete credibility features.

The other part of the research concerns conducting an experiment. The features extracted from the credibility framework are proposed to the crowd using surveys on a crowdsourcing platform.

The results of the crowdsourcing tasks are compared and analysed for similarities and differences. First, user comments are analysed for possible credibility factors, and similar factors are grouped. Next, features that strongly correlate ($r > 0.6$) with credibility are proposed to use in future research.

The proposed credibility features will not be used for further investigation, but are recommended for future research, for instance, to use the features as input for a credibility model or machine learning classifier.

1.1 Context

Whaling or whale hunting is targeted as the domain of this research, on the grounds that it is a controversial topic, and therefore probably contain a good balance between credible and non-credible news. Whaling events are closely related to activism and activism can lead to important social phenomena, like revolutions or terrorist attacks **[27]**. Since news can be spread quickly by anyone using the Internet, a chain reaction of activism can be triggered rather easily and eventually may lead to terrorism, for example, the Arab Spring[1] in 2010 and the rise of ISIS[2] in 2015. These examples stress the importance of the saliency of the news sources. Low-quality information may influence judgements of the crowd and form a distorted view of reality. Obtaining more inside in social phenomena may lead to

[1] https://en.wikipedia.org/wiki/Arab_Spring

[2] https://en.wikipedia.org/wiki/Islamic_State_of_Iraq_and_the_Levant

further arguments to support or not support social, ethical, or political causes. Analysing news sources and crowd opinions, interpreting them, and finding credibility features helps identifying justification or contradiction of activist events. Considering whaling is assumed to be a continuous topic of activism, the possibility to retrieve documents from a larger period of time is higher in contrast to topics of a temporal event, such as the Arab Spring or the rise of ISIS.

1.2 Problem Statement

In the present day there is an overwhelming abundance of information available on the web **[1, 12, 13, 18]**. The overabundance of that information caused many users to have issues managing and absorbing these information resources **[17]**. There have been many attempts in the present literature to overcome these issues with smart algorithms and applications to filter noise and retrieve both the relevant, reliable and novel stories **[14, 15, 16]**.

Present Twitter features and online news websites do not encompass a reliable and good method to find credible and relevant information on one specific event. For instance, if you search on an event term on Twitter you will only get a fire hose of tweets containing that term, most of the times completely irrelevant from the event. This problem is introduced by the advent of web democracy, the notion that everybody on the internet has the ability to publish content **[20]**. It would be helpful if the social network could get a comprehensive and dense timeline of relevant, important, and credible tweets. Discovering what factors relate to the credibility of the news may enhance the way the public looks at a tweet or article.

Traditionally, we relied on newspapers, television, magazines and similar media to gather the latest news. These traditional media contain content made primarily by journalists and media professionals. Because this group of professionals could easily be identified, independent committees could oversee them and ensure the neutrality of the news. Since the rise of social media, everyone was suddenly able to post "news" and share it with the world, without any moderation or consequences. It is assumed that therefore the factors indicating credibility of traditional media, such as online news articles, are different from the credibility factors of social media, such as tweets.

Recent studies **[11, 33]** investigated algorithms to calculate the credibility of tweets. The training sets consisted of tweets with a credibility assessment of the crowd, which were used to learn the importance of a pre-selected set of twitter features, with respect to the credibility of the tweet. Although this approach may help in the search of credibility of tweets, it was primarily focused on learning a credibility algorithm based on tweets with a rough credibility assessment (i.e. likely to be truth, definitely false). A research focused primarily on the features of the tweets is missing. Also, **[11]** indicates that the crowd experienced difficulty to assess the credibility of a tweet based on a score. Finally, an important missing gap might be the comparison between the credibility features of tweets and the credibility features of the online news articles.

1.3 Research Questions

This research aims at exploring the credibility measures for news sources from the web. The main research question is:

Q1 What is the perceived importance of credibility features in online news?

In order to answer the main research question, the following sub questions will be answered:

Q1.1 What is the perceived importance of credibility features in tweets?

Q1.2 What is the perceived importance of credibility features in online news articles?

Q1.3 What are the differences and similarities between the credibility features of tweets and online news articles?

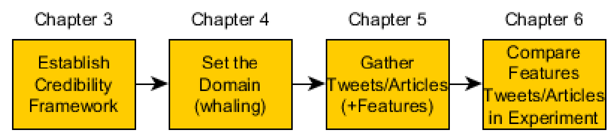

Figure 1: overview of the methodology

In Chapter 3 the literature is investigated with the aim of creating a theoretical framework around credibility. The credibility features described in Q1 are limited to the features resulted from this Credibility Framework.

Chapter 4 describes the domain of this research: whaling (whale hunting). The process of gathering events for this domain and the relevancy ranking method is explained.

Chapter 5 explains the data mining process and the properties of the gathered data.

In Chapter 6 the data is processed, so it fits the Credibility Framework. The methods and techniques applied in this process are also explained. Eventually, the result is the input for an experiment, which is used to find credibility features in tweets and news articles.

Chapter 7 analyses the results of the experiments. First, by looking at the results of the experiments individually. Later in this chapter, the credibility features of the tweets and news articles are compared.

Finally, in Chapter 8 the research questions are answered and discussed, and the research is concluded. The limitations of this research are reported and the future work that remains is suggested.

2. RELATED WORK

This section contains related work about automatic credibility determination for tweets and news articles on the web. Furthermore, the notion of crowdsourcing will be explained and two crowdsourcing platforms will be described. Finally, several methods of event detection in tweets and news articles will be explained.

2.1 Credibility in Online News

Credibility assessment in online news is a huge research topic for years. Compared to traditional media, such as newspapers and magazines, people tend to be more sceptic about the credibility of online news articles **[37]**. The ease of publishing news on the Internet is probably an important factor of this scepticism.

Online news can be divided in multiple groups, with respect to credibility. Examples are news articles, blogs, and social media. Within the social media group, Twitter belongs to the largest news medium, concerning to a recent study that indicated over 85% of the tweets in a large experiment were newsworthy **[38]**. In terms of credibility, people do not see it as more credible in contrast to blogs and news articles **[39]**. The credibility of blogs is highly dependent on the author. A survey among U.S. politically-interested Internet users showed that blogs sometimes can be seen as more credible in contrast to any online mainstream medium **[36]**.

Twitter, as a newsworthy social medium, produces an enormous amount of data **[14]**, which provides access to all kinds of news before it is published in the traditional news. Along with it the problem of the credibility of the tweets rises. Recent techniques like machine-learning help scientists to understand the credibility of the tweet in this fast changing digital world. A supervised learning approach in 2011 discovered that credible news can be indicated by the number of tweets of an author and the amount of retweets **[11]**. TweetCred is a browser plugin with a visual indication of the credibility of a tweet, using a credibility model based on more than 45 features **[40]**. This shows that theoretical models start to transform into concrete implementations, although such applications ask for continuous improvements.

2.2 Crowdsourcing Platforms

The rise of the Internet opened many doors and crowdsourcing[3] is one of them. Traditionally, people worked at a company to provide some kind of labour. Later, companies out-sourced the labour to companies in other countries to reduce costs. Today, the Internet provide us with platforms to post tasks, to be executed by "the crowd", anonymous workers that gain a small amount of money completing a micro task.

Nowadays, crowdsourcing can be found in many scientific projects. In computer science certain problems are really hard for a computer to solve, for instance, sentiment analysis and event detection **[41, 42]**. Taking in account the spammers among the crowd, crowdsourcing provides a way to gather a vast amount of robust data within limited time.

Amazon Mechanical Turk[4] (AMT) is a popular crowdsourcing platform providing the ability to create or execute micro tasks. Although AMT was a crowdsourcing pioneer, other platforms grow in popularity. For instance, Crowdflower[5] is a crowdsourcing platform with roughly the same abilities as AMT, but is extended with quality control, through gold question generation, and the ability to create reusable and configurable templates.

Researcher remain sceptic about the use of crowdsourcing platforms for academic purposes. Therefore, crowd evaluation platforms such as CrowdTruth **[44]** helps to discover spam and evaluate the quality of the results from the crowd. The CrowdTruth platforms aims at evaluating the disagreement among workers, which is the opposite of, for example, traditional agreement evaluation among domain experts.

3. CREDIBILITY

The definition of credibility can take many forms **[5, 9, 10, 35]**. In the domain of information sciences credibility is closely related to concepts as factuality, veracity, quality, trustworthiness, believability, sentiment, controversy and popularity **[4, 5, 6, 7, 8, 9, 10]**.

[3] https://en.wikipedia.org/wiki/Crowdsourcing

[4] https://www.mturk.com/mturk/welcome

[5] http://www.crowdflower.com

It may be worthwhile to derive concrete features from the abstract credibility definition, with respect to online news, because concrete features can be quantified and generalised. Furthermore, credibility features may differ depending on the context, which stresses the importance of investigating the body and metadata of online news for indicators of credibility. We developed the pyramid in Figure 2 which represents four levels of concreteness, starting from a very abstract layer on the top to a very concrete layer on the bottom. The more concrete a layer gets, the larger the number of credibility entities will be.

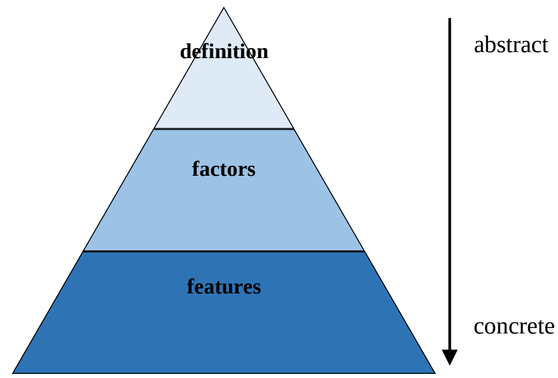

Figure 2: the credibility pyramid shows the different abstractions of credibility

The following paragraphs describe the different layers in the credibility pyramid, starting from the top, continuing towards the bottom.

3.1 Definition of credibility

We use the definition of credibility as stated in **[21]**, namely, "judgments made by a perceiver ... concerning the believability of a communicator". The table below describes the three concepts commonly used to elaborate on this credibility definition:

Table 1: credibility concepts used for elaboration on the definition of credibility

#	Concept	Description
c1	Competence	"The degree to which an observer believes that a target knows the truth" **[21]**
c2	Trustworthiness	"The degree to which an observer believes that a target will tell them the truth as he or she knows it" **[21]**
c3	Goodwill	"The degree to which a perceiver believes a sender has his or her best interests at heart" **[21]**

The perceiver in this context can also be a group, since mass media normally do not target an individual, but a target audience.

The definition of credibility provides us with the starting point in our search for credibility features in online news.

3.2 Factors of Credibility

We searched the literature for credibility factors, based on the credibility definition and its concepts. Credibility factors can be used to assess the credibility of tweets or articles using qualitative methods. Recent studies **[11, 22, 23, 24, 25, 26]** revealed numerous factors of credibility, in respect to online news. Table 2 lists the credibility factors, together with a description and reference to the literature. Additionally, we aligned each factor with its corresponding concepts (paragraph 3.1).

Table 2: Credibility factors

#	Factor	Description	Concept
f1	Commonality	"The more news publishers delivered articles with similar content to the target article being assessed, the higher the credibility was rated." **[22]**	Trustworthiness (c2) Goodwill (c3)
f2	Numerical Agreement	"Numerical expressions such as "100 passengers" or "three tracks" occur in news reports. When numerical expressions contradicted those in other articles from different news publishers, the credibility was rated lower." **[22]**	Competence (c1)
f3	Objectivity	"The credibility of articles containing subjective speculation was rated differently from those containing objective news sources." **[22]**	Trustworthiness (c2) Goodwill (c3)
f4	Spelling	The correctness of the spelling of the words in an article, as well as the grammar. **[23, 24]**	Competence (c1)
f5	Sentiment	The reactions that certain topics generate and the emotion conveyed by users discussing the topic: e.g. if they use opinion expressions that represent **positive** or **negative sentiments** about the topic.	Trustworthiness (c2) Goodwill (c3)
f6	Citations	The external sources cited: e.g. if they cite a specific URL with the information they are propagating, and if that source is a popular domain or not. **[11]**	Trustworthiness (c2) Goodwill (c3)
f7	User activity	"The extent to which an author is active on a news platform. For example, the number of posts or the number of comments." **[11]**	Competence (c1)
f8	News source	The source of an article (e.g. Guardian or Twitter) **[26]**	Competence (c1)
f9	Level of expertise	The level of expertise of an author. The expertise dimension includes for example quality, accuracy, authority, and competence. **[25]**	Competence (c1)

3.3 Credibility Features in Online News

Although the factors of credibility tend to be concrete, they are not always measurable in a quantitative way. Furthermore, the factors cannot always directly be aligned to features of a tweet or news article.

In this research we aim at finding credibility features within the textual content and metadata of tweets and articles. Looking at the features used in previous studies **[11, 33, 39]** about credibility in online news, we can group features into multiple categories, which we call feature dimensions. The credibility feature dimensions are listed in Table 3. We believe that each dimension has its own scope and therefore it might be interesting to group tweet or article features per feature dimension.

Table 3: Credibility feature dimensions

#	Dimension	Description
d1	Author	The author of the article or tweet and its characteristics. For example, the number of followers or the number of posts.
d2	Content	The content of the article or tweet and its characteristics. For example, grammar faults, number of nouns, number of words.
d3	Interaction	The interaction from other people with the article or tweet and its characteristics. For example, number of comments or retweets or the number of favourites.
d4	Sentiment	The level of emotions expressed in the article or tweet. For example, the use of emoticons or words that express sentiment.

This paragraph provides Table 4 and Table 5 (credibility features) with tweet and article credibility features, aligned with the credibility factors from Table 2. Every feature is assigned to a feature dimension from Table 3.

By default the results retrieved from the used API's (Chapter 5) come with a vast amount of features. We assessed these features for credibility based on the credibility factors of Paragraph 3.2. Additional features are computed to cover all credibility factors and to cover the feature dimensions. Unfortunately, the "Numerical agreement" and "Objectivity" factors are not covered for tweets and articles, because no standard techniques or of-the-shelf API's are offered to gather features based on these factors. Also, the "News Source" factor is not covered, because the tweets all come from the same source (Twitter) and the articles all originate from either The Guardian or The New York Times. The API's for the news articles do not provide features in the results related to "Citations" or "Level of Expertise" factors. Furthermore, the API's for the articles did only provide the names of the authors of an article, which made it insufficient to find features for the Author dimension.

3.3.1 Credibility features for tweets

By default the tweets retrieved from the Twitter API (Chapter 5) come with a rich set of properties of the tweet, for example retweets count, author information and information about the entities in the tweet. The properties that relate to one of the credibility factors (Table 2) are extracted and added as a credibility feature. Additional features are populated using natural language processing, sentiment analysis and off-the-shelf API's. For example, the sentiment140 API aims at sentiment analysis specifically for tweets. Newser.com is an online news aggregator with a ranked list of references to popular online news sources.

This list is used as input to test the influence of news site popularity from an external source. Another way to find external references concerns the usage of an API that analyses a text for terms referring to a Wikipedia page.

Table 4: Credibility features for tweets

Feature	Description	Factor	Dimension
created_at	Date the tweet was created	User activity (f7)	Content (d2)
favorite_count	number of likes	Commonality (f1)	Interaction (d3)
retweet_count	number of retweets	Commonality (f1)	Interaction (d3)
word_count	Number of words	Commonality (f1)	Content (d2)
symbol_occurrences	Number of symbol occurrences	Spelling (f4)	Content (d2)
uppercase_count	number of uppercase characters	Spelling (f4)	Content (d2)
numerical_occurrences	number of numerical occurrences (e.g. 100, thousand)	Commonality (f1)	Content (d2)
wikipedia_entity_count	number of official wikipedia entities in text	Spelling (f4) Citations (f6)	Content (d2)
proper_nouns	number of proper nouns in text	Spelling (f4)	Content (d2)
mentions_count	number of mentions in tweet	Citations (f6)	Content (d2)
url_count	number of URL's in tweet	Citations (f6)	Content (d2)
hashtag_count	number of hashtags in tweet	Citations (f6)	Content (d2)
url_in_newser100	contains a URL from the Newser.com top 100	Commonality (f1) Citations (f6)	Content (d2)
url_in_newser	contains a URL from listed on Newser.com	Commonality (f1) Citations (f6)	Content (d2)
user_friends_count	number of friends	User activity (f7)	Author (d1)
user_listed_count	number of friend lists	User activity (f7)	Author (d1)
user_followers_count	number of followers	User activity (f7)	Author (d1)
user_has_url	profile contains a URL	User activity (f7)	Author (d1)
user_has_description	length of profile description	User activity (f7)	Author (d1)
user_created_at	the date the profile was created	User activity (f7)	Author (d1)
user_verified	the account is officially verified by Twitter	Level of expertise (f9)	Author (d1)
user_has_profile_image	user uses custom profile image	User activity (f7)	Author (d1)
user_has_custom_background	user uses custom profile background	User activity (f7)	Author (d1)
sentiwordnet	sentiment analysis based on the sentiwordnet corpus	Sentiment (f5)	Sentiment (d4)
sentiment140	sentiment analysis based on the sentiment140 corpus	Sentiment (f5)	Sentiment (d4)

3.3.2 Feature extraction articles

By default news articles retrieved from the news article API's (Chapter 5) come with a set of properties about the article, for example date published, author information and keywords. The properties we think are relevant to one of the credibility factors (Table 2) are extracted and added as a credibility feature. Additional features can be populated using natural language processing or sentiment analysis.

Table 5: Credibility features for articles

Feature	Description	Factor	Dimension
published_date	The date the article was published	User activity (f7)	Content (d2)
uppercase_count	number of uppercase characters	Spelling (f4)	Content (d2)
proper_nouns	number of proper nouns in text	Spelling (f4)	Content (d2)
comments_count	number of comments in article	Commonality (f1)	Interaction (d3)
keywords_count	number of keywords in article	Commonality (f1)	Content (d2)
numerical_occurrences	number of numerical occurrences (e.g. 100, thousand)	Commonality (f1)	Content (d2)

dbpedia_entity_count	number of official dbpedia entities in text	Spelling (f4)	Content (d2)
symbol_occurrences	Number of symbol occurrences	Spelling (f4)	Content (d2)
word_count	Number of words	Commonality (f1)	Content (d2)
sentiwordnet	sentiment score	Sentiment (f5)	Sentiment (d4)

4. EVENT SPACE

Since the domain of this research consists of the "whaling" or "whale hunting" activist event, as described in the context (paragraph 1.1), some constraints are used to limit the datasets to a certain scope. Within this research the scope created around the whaling event is called the event space.

Domain experts from the social science department provided us with a collection of seed words (appendix I) to create a general sense about the whaling event. The seed words are used to mine tweets and articles. The data mining process and data set characteristics are described in Chapter 5.

The event relevancy is measured by calculating the coverage of the seed words in the article or tweet. The event relevancy score consists of the total number of matching seed words in the tweet or article, divided by the total number of seed words.

event relevancy = matched seed words / total seed words

Sorting the tweets or articles descending, based on the event relevancy score, results in a list with the most relevant tweets in the top and less relevant articles in the bottom. We use the top of this ranked list as input data for our experimental design.

5. DATA

In this research we will use the datasets as presented in table 5. The datasets are related to the "whaling" event as described in chapter 4.

Table 6: Datasets, including the type, period, source and size

#	Type	Period	Source	Size
DS1	Tweets	March 2015 - May 2015	Twitter Streaming API[6]	24150
DS2	News articles Blogs Opinion pages	January 2010 - June 2015	NYTimes articles API[7] NYTimes Community API[8] Guardian Open Platform[9] www.guardian.com www.nytimes.com	227

The tweets (DS1) are mined from the Twitter platform using the Twitter Streaming API, using roughly the seed words provided in appendix I, although some seed words are translated into hashtags and search keywords. To avoid getting too much spam, some hashtags are provided with additional keywords, such as "whaling" or "whale". Other seed words are left out intentionally, because they are too trivial. For instance, "Japan" and "quota" will result in a huge amount of tweets not related to the whaling event. The tweets were gathered in the spring of 2015 and the dataset contains a gap between the end of March and the beginning of April, because of problems with the API.

The second dataset (DS2) contains news articles, blogs, and opinion pages, gathered using the New York Times articles and community API and the Guardian Open Platform API, using the seed words provided in appendix I. Blogs and opinion pages may contain more controversial and subjective news, which may help to provide a better balance in the dataset. Furthermore, information like the body of the article and the comments have been scraped from nytimes.com and guardian.com using the Microdata entities in the source code of the online articles. The news articles consist of World and U.S. news with a material type of "brief" or "news". The blogs and opinion pages consists of the "opinion" section of the New York Times and the articles with a material type of "blog". The blogs from the Guardian are filtered based on the tag "tone/blog".

The next chapter describes how these data sets are analysed, enriched and evaluated.

[6] https://dev.twitter.com/streaming/overview/

[7] http://developer.nytimes.com/docs/read/article_search_api_v2/

[8] http://developer.nytimes.com/docs/community_api/The_Community_API_v3/

[9] http://open-platform.theguardian.com/

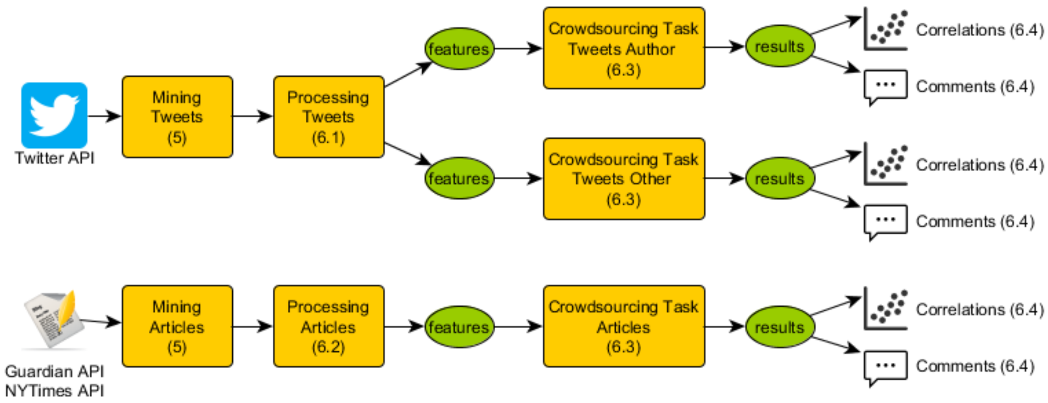

Figure 3: This diagram describes a simplified overview of the data processing pipeline. It consists of two paths: the tweet and article path. The yellow boxes and the result images on the right are aligned with the paragraphs of chapter 5 and 6.

6. PIPELINE

This chapter describes the methods and techniques used to analyse and process the datasets DS1 and DS2. Figure 3 offers a simplified overview of the data processing pipeline and can be used as a reference throughout this chapter. The repository of the source code used to retrieve and process the tweets and articles is available on Github[10].

The pipeline for processing the datasets DS1 (tweets) and DS2 (articles) may look similar, but differ at several core points. Therefore, two different descriptions for each pipeline will be presented in this chapter. First the datasets will be processed, resulting in three CSV files with input data for the crowdsourcing tasks. The next step is the development of crowdsourcing templates. The input data and templates result in three crowdsourcing surveys containing the credibility features from the previous step. Finally, the methods to analyse the output data from the crowdsourcing experiment is described in the last paragraph.

6.1 Processing Tweets

The tweets from dataset DS1 contain the credibility features as described in Paragraph 3.3.1. Some of this features are selected directly from the data retrieved from the Twitter API, other features are extracted by analysing the content of the tweet.

6.1.1 Feature selection

By default, the Twitter API returns advanced result sets with a vast amount of features. Only features related to credibility are extracted, where the selection is based on the credibility factors in table 2:

date created, #favourites, #retweets, #mentions, #URL's, #hashtags, #friends, #friend lists , #followers, profile contains a URL, length profile description, date user created, user verified by Twitter, user uses custom profile image, user uses custom profile background

The features representing a count will be counted using the *len* function of python. Some Boolean functions (like user uses custom profile background) check the tweet or article on the existence of certain elements, which results in either true or false.

[10] https://github.com/mvjacobs/Crone/

6.1.2 Feature extraction

Other features are propagated using natural language processing (NLP) and sentiment analysis. The table below describes the extracted features and the tools that will be used to propagate them:

Table 7: Tools, libraries and techniques used for feature extraction of tweets

feature	tools
#words	Regular expression (\S+)
#symbols	TweeboParser **[28]**, Twokenize **[29]**
#uppercase characters	Isupper function of Python
#numerical occurrences	TweeboParser **[28]**, Twokenize **[29]**
#wikipedia entities	Tagme[11] API
#proper nouns	TweeboParser **[28]**, Twokenize **[29]**
Contains newser top 100 url	The top 100 list of newser.com[12]
Contains newser url	All URL's from newser.com
Sentiwordnet score	NLTK with sentiwordnet corpus
Sentiment140 score	Sentiment140[13] API

6.1.3 Output format

After the feature extraction and selection the tweet result set consist of credibility features and the text of the tweet. We take the top 200 tweets, ranked on event relevancy. The input for the crowdsourcing platform we use takes a comma separated file (CSV) as input for a task, so we create the CSV with the following headers:

- *Tweet1* ID, *Tweet1* Text, *Tweet1* Feature1, *Tweet1* Feature2, … , *Tweet1* FeatureN;
- *Tweet2* ID, *Tweet2* Text, *Tweet2* Feature1, *Tweet2* Feature2, … , *Tweet2* FeatureN.

This means that one row in the CSV files contains the information of two tweets, on the grounds that putting all information in one row for one question in the crowdsourcing task makes it easier to create the template (Paragraph 6.3.2).

[11] http://tagme.di.unipi.it/

[12] http://www.newser.com/topsites.aspx

[13] http://www.sentiment140.com/

The tweets are added to the CSV file as follows: tweet 1 and tweet 2 on one row, tweet 2 and tweet 3 on one row, and eventually tweet 200 and tweet 1 on one row. Finally, the CSV consists of 101 rows, including the header. Each tweet is compared to two different tweets (to its left and right neighbour), because a pairwise comparison between all tweets would make either the task too large or we could only use a few number of tweets.

We assume that putting all 26 features in one crowdsourcing task will make the judgement of the crowd workers less accurate, because analysing all these features every question is a time-consuming task. Therefore, we split the features based on their dimensions (Paragraph 3.3.1). We take the largest feature dimension (author) and create a separate task, which results in an "author" task with 10 features and an "other" task (containing the features of the other three dimensions) with 16 features. The input CSV's for the two tweet tasks is eventually distributed as follows:

- CSV1 contains author-related features from the author (d1) dimension;
- CSV2 contains the rest of the features from the content (d2), interaction (d3) and sentiment (d4) dimensions.

This also splits the tweet task into two tasks, which will be further explained in paragraph 6.3.

6.2 Processing Articles

The articles from dataset DS2 contain credibility features as described in paragraph 3.4.2. Some of these features are selected directly from the data retrieved from the New York Times API or Guardian Open Platform, other features are extracted by analysing the content of the article.

6.2.1 Article normalisation

News articles, blogs and opinion pages are retrieved from the New York Times API and The Guardian Open Platform. By default, querying the API produces results in different formats. Therefore, the results will be compared and similar features will be mapped, which will result in one normalised result set (appendix II).

6.2.2 Feature selection

Data mining using New York Times API and The Guardian Open Platform will result in several out-of-the-box features.

title, #words, source, url, author, keywords, publication date, article, abstract, comments, #comments

The features representing a count are count using the *len* function of python. The text for the article and comments for the Guardian and the comments for the New York Times are extracted using the Microdata[14] available on the page of the article, following the conventions of schema.org. When the abstract field is empty, the abstract consists of the first two sentences of the article, extracted by the sentence tokenizer[15] of the NLTK library from Python.

6.2.3 Feature extraction

Other features will be propagated using natural language processing (NLP) and sentiment analysis. Table 8 describes the complete list of the computed article credibility features.

[14] http://www.w3.org/TR/microdata/

[15] http://www.nltk.org/_modules/nltk/tokenize/punkt.html

Table 8: Tools, libraries and techniques used for feature extraction of articles

feature	tools
#symbols	NLTK, Part-of-speech entities: [SYM] [,] [.] [:] [#] [$] [)] [(]
#uppercase characters	Isupper function of Python
#numerical occurrences	NLTK, Part-of-speech entities: [CD]
#wikipedia entities	Tagme[16] API
#proper nouns	NLTK, Part-of-speech entities: [NNP] [NNPS]
Sentiwordnet score	NLTK with sentiwordnet corpus

6.2.4 Output format

After the normalisation, feature extraction and selection, the articles consist of credibility features and the abstract of the article. The structure of the CSV file, used as input data for the crowdsourcing task, is the same as the CSV files of the tweets in Paragraph 6.1.3, except for the fact that the Tweet Text header is replaced with the Article Abstract. The input CSV for the article task (CSV3) is therefore structured as follows:

- *Article1* ID, *Article1* Abstract, *Article1* Feature1, *Article1* Feature2, … , *Article1* FeatureN;
- *Article2* ID, *Article2* Abstract, *Article2* Feature1, *Article2* Feature2, … , *Article2* FeatureN;

6.3 Experimental Design

The processed tweets and articles (Paragraph 6.2) are stored in the following files:

- 1x CSV of 100 tweet combinations (author features)
- 1x CSV of 100 tweet combinations (other features)
- 1x CSV of 100 article combinations

In this paragraph we describe the three crowdsourcing tasks we create with the three CSV files from above as input. Also, the template for the crowdsourcing task is described, and the tools and techniques for the template creation are explained.

6.3.1 Tasks

The three CSV files from above are used to create three crowdsourcing tasks: two for tweets and one for articles.

Table 9: overview of crowdsourcing tasks

#	Description	Input data
T1	Compare tweets with author features	CSV1
T2	Compare tweets with other features	CSV2
T3	Compare articles	CSV3

The crowd consists of people from the United States, United Kingdom, Australia, and Canada. Every question is answered by 10 workers and a worker will earn 0.02 dollar per question.

The crowd worker (or crowd user) compares two tweets or articles and the corresponding features. Next, the worker determines which tweet or article is more credible. A third option is to determine that the tweets or articles are equally credible. Every

[16] http://tagme.di.unipi.it/

chosen option changes the headings of the tweets to a different colour, to emphasise which tweet is selected to be more credible. The following step consists of clicking the features that may determine the credibility. Optionally, a comments field is provided, for instance, to give additional information about the choices the worker made. For anti-spam purposes two constrains validate the question form: an answer on the credibility question is required to continue to the next question, and selecting no features will produce a warning. The warning informs the worker to either continue to the next question, or to provide additional features for the current question. Nonetheless, providing no features can also be an option.

6.3.2 Template

The crowdsourcing platform we use is Crowdflower[17], since the creation of templates for the tasks is highly customisable.

[17] http://www.crowdflower.com

Therefore, the template for the tasks consists of CML, the scripting language of Crowdflower, in combination with HTML5, CSS, and Javascript. Additionally, we take advantage of the Twitter Bootstrap[18] front-end framework, which provides tools for a professional look-and-feel.

The main layout of the template consists of the following elements:

- Explanation about the task;
- Two columns with the tweets/articles and the corresponding features;
- Credibility question with possible answers as radio buttons;
- Comments field.

Figure 4 contains a visualisation of the user interface and the explanation presented to the worker. In Appendix III all templates are provided.

[18] http://getbootstrap.com/

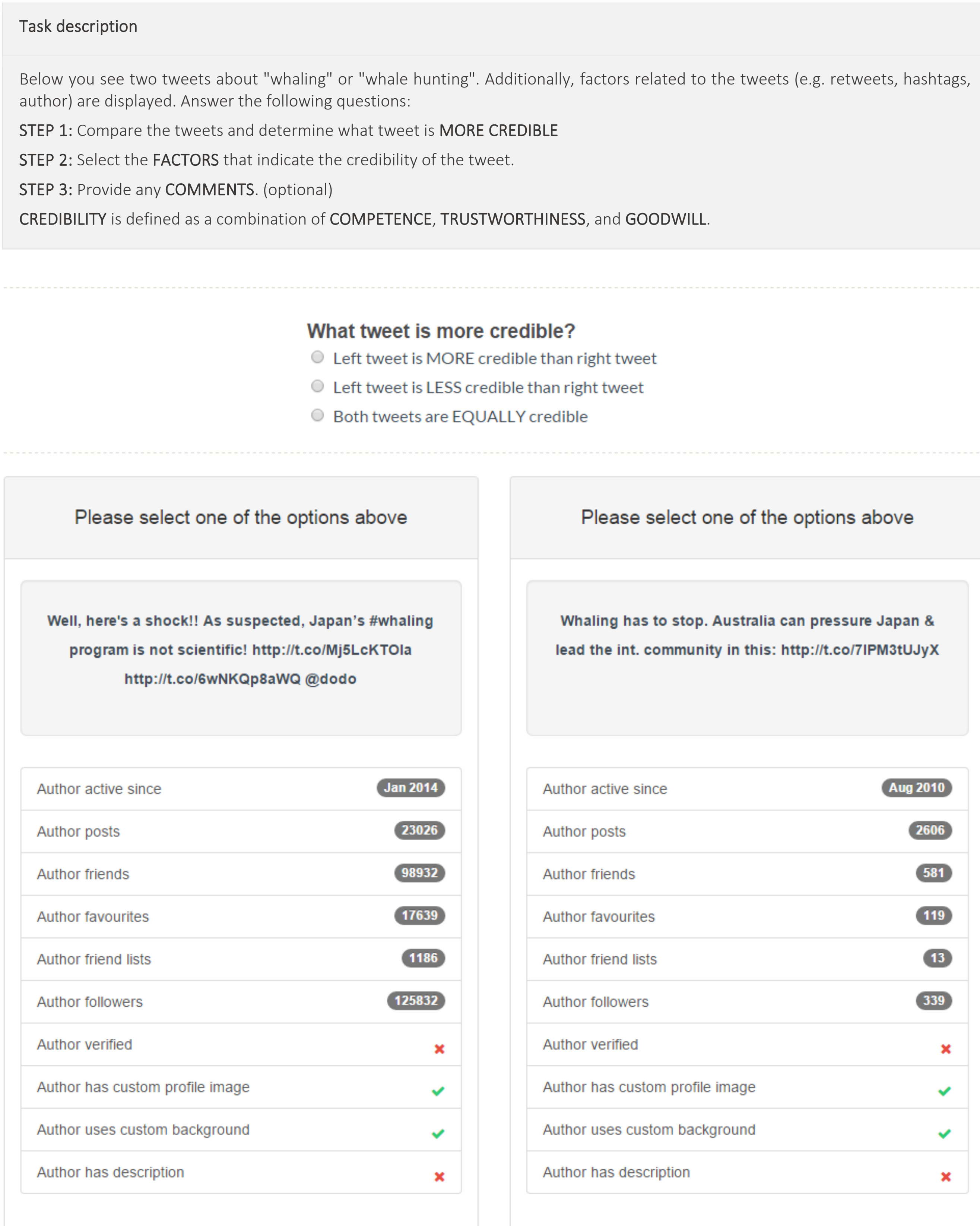

Figure 4: the CrowdFlower user interface of the "tweets with author features" experiment. The text area used to provide additional comments is located below the tweet feature panels (not visible in this screenshot).

6.4 Results analysis

After the workers completed the crowdsourcing tasks three result sets are available on the CrowdFlower platform. The results are exported as CSV files and manually checked for unexplainable content, distributed in repetitive patterns, and this spam will be removed.

We collect and analyse the comments. Hence, the comments are searched for credibility features and factors, which are categorised in a coding table (Appendix IV).

Next, the results are analysed in a quantitative way by determining the level of correlation between the times the tweet itself is marked as more credible than the other tweet and the times a feature is marked as an indicator for the credibility. We will visualize the correlation charts per feature and calculate the Pearson correlation coefficient. For this analysis we only use the tweets or articles that are marked as more credible, together with the corresponding features which may indicate the credibility. The charts are generated in R[19].

The outcome of the analysis is described in the results chapter.

7. RESULTS

The crowdsourcing tasks, as described in Paragraph 6.3.1, were finished within three days. Table 10 describes the statistics of the results of the three tasks.

Table 10: #workers, comments and results of the three tasks

Task	#Workers	#Comments	#Results
Tweets Author (T1)	131	122	1000
Tweets Other (T2)	122	101	1000
Articles (T3)	136	111	1000

Two areas of the results retrieved from the crowdsourcing tasks are described in this chapter: the analysis of the comments of the workers (Paragraph 7.1), and the correlation between the number of tweets indicated as more credible and the count of each feature indicated as credible feature (Paragraph 7.2).

7.1 Crowd Worker Comments

Appendix IV describes the comments extracted from the three crowdsourcing tasks. Each comment is analysed using a qualitative coding method. First, words, word spans, or sentences indicating credibility are reduced to one or two words. Next, we analyse the words for similarities and group similar words into categories. We qualify the category names as credibility factors.

The credibility factors extracted from the comments of the tweets with author features are:

Human-like stats (5), Author friends count (4), Evidence (4), Objectivity (2), Symbolic occurrences (2), Clarity (2), Style of the tweet (2), Factuality, Completeness, Author followers count, Authority, Post count too high, Contains hyperlink, Controversy, Author friends count, Written numbers, Sentiment, Focus, Combination of profile customisation, author followers count, post count, friends count

The credibility factors extracted from the comments of the tweets with other features are:

Sentiment (5), Objectivity (2), Too many hashtags, Factuality, Sustainability, Sarcasm, Spelling/grammar

The credibility factors extracted from the comments of the articles are:

Completeness (4), Tone (2), Objectivity (2), Level of detail (2), Novelty (2), Factuality, Brevity, Clarity, Multiple viewpoints, Fewer Wikipedia entities, Content of the comments, Importance, Quotes in article, Numerical occurrences, References

Furthermore, one worker declared that the features presented in the tweets do not represent credibility at all. Another worker pointed out that counting characters and counting words have nothing to do with credibility.

7.2 Correlation Analysis

Each result set is analysed for correlation between the "times a tweet or article is marked as more credible" and the "times a feature is indicated as credible". Therefore, scatter plots are created (Appendix V) and the Pearson's correlation coefficients are calculated for every credibility feature. We used the following categories for indicating the strength of the correlation:

Table 11: correlation categories

Correlation strength	Correlation range
Weak positive correlation	$r < 0.3$
Light positive correlation	$r > 0.3$ and $r < 0.6$
Strong positive correlation	$r > 0.6$
Weak negative correlation	$r < -0.3$
Light negative correlation	$r > -0.3$ and $r < -0.6$
Strong negative correlation	$r > -0.6$

The results of each crowdsourcing task are displayed in a separate table.

Table 12: correlation analysis of tweet author features

Feature	r	Strength
Author has avatar	0.3209515	Light +
Author has background	0.1256566	Weak +
Author has description	0.09985135	Weak +
Author registration date	0.3936765	Light +
Author #favourites	0.4263987	Light +
Author #followers	0.7077715	Strong +
Author #friends	0.5740222	Light +
Author #friend lists	0.3188064	Light +
Author #posts	0.5993252	Light +
Author is verified	0.03866856	Weak +

Table 13: correlation analysis of tweet "other" features

Feature	r	Strength
Sentiment140 score	0.33669904	Light +
Sentiwordnet score	0.21392628	Weak +
Created date	0.37941613	Light +

[19] https://www.r-project.org/

#favourites	0.23966669	Light +
#hashtags	0.27399205	Light +
#nouns	0.31915700	Light +
#numerical entities	-0.01002103	Weak -
#symbols	0.08215554	Weak +
#retweets	0.52850012	Light +
#uppercase characters	0.51144734	Light +
#urls	0.26363402	Weak +
#user mentions	0.16022874	Weak +
#wikipedia entities	0.60237539	Strong +
#words	0.53088918	Light +
URL in Newser list	0.27691787	Weak +
URL in Newser top 100 list	0.19338630	Weak +

Table 14: correlation analysis article features

Feature	r	Strength
#comments	0.06630054	Weak +
#keywords	0.22329140	Weak +
#nouns	0.34554077	Light +
#numerical entities	0.23826949	Weak +
Publication date	0.20255478	Weak +
#symbols	0.12981450	Weak +
#uppercase characters	0.23487976	Weak +
#wikipedia entities	0.48243573	Light +
#words	0.40601756	Light +
Sentiment score	0.36201947	Weak +

In Table 15 the correlation differences between corresponding features in tweets and articles are compared. If the difference between correlation values (r) is larger than 0.2, we consider a feature as "different".

Table 15: credibility features: tweets vs articles

Feature	Tweet	Article	difference in r > 0.2
#retweets/#comments	0.52850012	0.06630054	yes
#hashtags/#keywords	0.27399205	0.22329140	no
#nouns	0.31915700	0.34554077	no
#numerical entities	-0.01002103	0.23826949	yes
Publication date	0.37941613	0.20255478	no
#symbols	0.08215554	0.12981450	no
#uppercase characters	0.51144734	0.23487976	yes
#wikipedia entities	0.60237539	0.48243573	no
#words	0.53088918	0.40601756	no
Sentiment score	0.21392628	0.36201947	no

8. DISCUSSION

During this research three crowdsourcing tasks have been conducted, where two tasks were focussed on the credibility features of tweets and one on the credibility features of online news articles. The crowd workers have been asked to compare two tweets or articles and select the one that seems the most credible, or to mark them both as equally credible. Additionally, a set of credibility features have been presented where the crowd worker selected the features that seemed the most important to determine the credibility of the tweet or article.

Q1 What is the perceived importance of credibility features in online news?

A1 In tweets *followers count* (Table 12) and *Wikipedia entity count* (Table 13) are strongly correlated credibility features ($r > 0.6$). The features for the news articles did not have a strong correlation.

Crowd workers perceived *Human-like stats* and *sentiment* (mentioned by five different worker) as important credibility factors for tweets. *Completeness* (mentioned by four different workers) is perceived as an important credibility factor for news articles.

The answer on the main research question is based on the answers of the sub questions, which are answered in the following paragraphs.

Q1.1 What is the perceived importance of credibility features in tweets?

A1.1 In tweets *followers count* (Table 12) and *Wikipedia entity count* (Table 13) are strongly correlated credibility features ($r > 0.6$).

The results of the correlation analysis between the "tweets or articles marked as more credible" and the "features that have been selected as indicators of credibility" showed that the highly correlated feature ($r > 0.6$) for the "author of a tweet" task was the *followers count* of the author of a tweet. The number of followers may aim at the popularity of the author. The *Wikipedia entity count* in a tweet is highly correlated ($r > 0.6$) in the "other tweet features" task. The Wikipedia entity count may refer to entities that exist in the real world, which likely increases the truthfulness of the named entities in the tweet.

Numerical author features are lightly correlated ($r < 0.6$ and $r > 0.3$) and stronger correlated than the Boolean features of the author, which are all weakly correlated ($r < 0.3$), except the *author has avatar* feature.

The analysis of the comments of the crowd workers stated *Human-like stats* and *sentiment* (mentioned by five different crowd workers) as important credibility factors for tweets.

Q1.2 What is the perceived importance of credibility features in online news articles?

A1.2 No strongly correlated ($r > 0.6$) features are present for the news articles. *Wikipedia entity count* ($r = 0.48$) and *word count* ($r = 0.40$) are the strongest correlated features for news articles.

The credibility assessment for articles may be harder than for tweets, because of the difference in size of the content, but also because there are limited features available. For example, details about the author of the article were not available through the API.

However, it is interesting that Wikipedia entity count has the highest correlation score for news articles. This may, just as for the tweets, indicate the amount of Wikipedia entities in a text helps to determine the credibility of a news item.

The analysis of the comments of the crowd workers stated *Completeness* as an important credibility factor for news articles (mentioned by four different crowd workers). The number two correlated feature *word count* (r=0.4) may relate to this factor, since a more complete article often means more content as well.

Q1.3 What are the differences and similarities between the credibility features of tweets and online news articles?

A1.3 The features *#retweets/#comments* and *#uppercase characters* are stronger correlated (difference in r > 2) with credibility in tweets than in news articles. On the other hand, *#numerical entities* is stronger correlated in news articles than in tweets. The correlation difference for the other features in Table 15 is similar, comparing tweet features with article features.

The correlation for the *number of comments* for an article is rather low (r < 0.1). The correlation for the *retweet count* feature for tweets is quite high (r > 0.6). This is an important difference comparing credibility features in tweets versus credibility features in news articles, considering a retweet as a comment of a tweet. This may point out that the credibility importance of social activity on news sites, opinion pages, or blogs is lower than the social activity on Twitter. Another difference between tweets and articles is the *uppercase character count*. The uppercase characters in tweets may refer to "yelling" in tweets, which probably does not happen often on newspaper sites.

On the other hand, *Wikipedia entity count* and *word count* have similar correlation scores. The word count may refer to the completeness of the articles and tweets. The Wikipedia entity count may refer to amount of referencing to real world entities.

8.1 Limitations

The absence of retweets and comments is an important limitation of this research. The whaling topic lacks popularity and therefore not many people retweet or comment on the articles. In future research a more popular news event can be chosen, or an experiment can be conducted that uses a data set with features consisting of values that are equally distributed.

Although the ranked event list provided a tweets or articles highly related to the whaling event, not all spammers could be avoided. More comprehensive machine-based spam filtering may result in a list with more human-like statistics.

Not all news articles contained abstracts. Therefore, the first two sentences of the text are used as an abstract. It is likely that these sentences did not contain all the important information of the article. Summary generators for creating abstract could be a solution for this problem.

In this research we used a small set of features, covering a large part of the feature dimensions and credibility factors. For example, no author features were available for the news articles, although this may be an important part of credibility assessment of an article.

Another limitation is the amount of data. Expanding the amount of tweet or article comparisons may result in more accurate results.

The source to gather the articles from were limited to the New York Times and The Guardian. Adding more sources may create a more diverse sample, which me be more representative for the task.

Not all tweets and articles retrieved were newsworthy. Tweets promoting petition signing and chatter slipped through, which points out that the event relevancy ranking may not be enough to gather the relevant data.

8.2 Future work

The highly correlated features are suggested to use in credibility algorithms or models, which can also be expanded with related features related to the credibility factors extracted from the worker comments.

We suggest the use of a larger dataset in future research. This most likely solves the problem of unequally distributed values of the credibility features.

Another suggestion is the use of the pairwise comparison method to completely compare all tweets pairs and articles. This will result in a better representation of reality.

An interesting task, which was not part of the scope of this research, is the analysis of the values of the credibility features that are marked as credibility indicators. This may result in a credibility range for the specified feature.

Combining multiple features into one may result in an increased credibility importance. For example, combining number of posts, retweets, and favourites may provide a factor to evaluate human-like statistics for an author.

Analysing features from referenced sources, mentioned people, or comments or retweets, may influence the credibility of a tweet or article. For instance, a tweet mentions a controversial organisation can decrease the credibility of that tweet.

Finally, sentiment analysis is a popular topic of research these days **[32, 33, 34]**. The use of more comprehensive sentiment corpora helps detecting controversial or biased tweets and articles, which possibly relate to either a low credibility of the content or the author.

REFERENCES

[1] Vossen, P. H. (2012). The Challenge of Information Balance in the Age of Affluent Communication. Information Overload: An International Challenge for Professional Engineers and Technical Communicators, 41-59.

[2] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 55-60).

[3] Ploeger, T., Kruijt, M., Aroyo, L., de Bakker, F., Hellsten, I., Fokkens, A., Hoeksema J. & ter Braake, S. (2013). Extracting activist events from news articles using existing NLP tools and services. Detection, Representation, and Exploitation of Events in the Semantic Web, 30.

[4] Waguih, D. A., & Berti-Equille, L. (2014). Truth Discovery Algorithms: An Experimental Evaluation. arXiv preprint arXiv:1409.6428.

[5] Nagura, R., Seki, Y., Kando, N., & Aono, M. (2006, August). A method of rating the credibility of news documents on the web. In Proceedings of the 29th

annual international ACM SIGIR conference on Research and development in information retrieval (pp. 683-684). ACM.

[6] Wöhner, T., & Peters, R. (2009, October). Assessing the quality of Wikipedia articles with lifecycle based metrics. In Proceedings of the 5th International Symposium on Wikis and Open Collaboration (p. 16). ACM.

[7] Rieh, S. Y., Jeon, G. Y., Yang, J. Y., & Lampe, C. (2014). Audience-aware credibility: From understanding audience to establishing credible blogs.

[8] Vydiswaran, V. G., Zhai, C., & Roth, D. (2011). Content-driven trust propagation framework. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 974-982). ACM.

[9] Kohring, M., & Matthes, J. (2007). Trust in news media development and validation of a multidimensional scale. Communication Research, 34(2), 231-252.

[10] Johnson, T. J., & Kaye, B. K. (2009). In blog we trust? Deciphering credibility of components of the internet among politically interested internet users. Computers in Human Behavior, 25(1), 175-182.

[11] Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In Proceedings of the 20th international conference on World wide web (pp. 675-684). ACM.

[12] Sankaranarayanan, Jagan, et al. "Twitterstand: news in tweets." Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems. ACM, 2009.

[13] Our mission: To give everyone the power to create and share ideas and information instantly, without barriers: 2015. https://about.twitter.com/company Accessed: 2015-31-05.

[14] Gabrilovich, Evgeniy, Susan Dumais, and Eric Horvitz. "Newsjunkie: providing personalized newsfeeds via analysis of information novelty." Proceedings of the 13th international conference on World Wide Web. ACM, 2004.

[15] Chu, Wei, and Seung-Taek Park. "Personalized recommendation on dynamic content using predictive bilinear models." Proceedings of the 18th international conference on World wide web. ACM, 2009.

[16] Yang, Yiming, et al. "Topic-conditioned novelty detection." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[17] Del Corso, Gianna M., Antonio Gullí, and Francesco Romani. "Ranking a stream of news." Proceedings of the 14th international conference on World Wide Web. ACM, 2005.

[18] Internet in Real-Time: How Quickly Data is Generated: 2015. http://pennystocks.la/internet-in-real-time/: Accessed: 2015-31-05.

[19] Ferrucci, David A. "Introduction to “this is watson”." IBM Journal of Research and Development 56(3.4). pages 1-1. 2012.

[20] Descy, Don E. "The Wiki: True Web Democracy." TechTrends 50.1, pages 4-5. 2006

[21] Westerman, D., Spence, P. R., & Van Der Heide, B. (2014). Social Media as Information Source: Recency of Updates and Credibility of Information. Journal of Computer-Mediated Communication, 19, 171–183.

[22] Nagura, R., Seki, Y., Kando, N., & Aono, M. (2006). A Method of Rating the Credibility of News Documents on the Web. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 683–684.

[23] Schwarz, J., & Morris, M. (2011). Augmenting web pages and search results to support credibility assessment. Proceedings of the 29th SIGCHI Conference on Human Factors in Computing Systems, 1245–1254.

[24] Gün, A., & Karagöz, P. (2014). A hybrid approach for credibility detection in Twitter. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8480 LNAI, 515–526.

[25] Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., … Hall, C. (2001). What Makes Web Sites Credible ? A Report on a Large Quantitative Study CHI 2001. Sigchi’01, 61–68.

[26] Schmierbach, M., & Oeldorf-Hirsch, A. (2012). A Little Bird Told Me, So I Didn’t Believe It: Twitter, Credibility, and Issue Perceptions. Communication Quarterly, 60(January 2015), 317–337.

[27] Ploeger, T., Armenta, B., Aroyo, L., De Bakker, F., & Hellsten, I. (2012). Making sense of the arab revolution and occupy: Visual analytics to understand events. In CEUR Workshop Proceedings (Vol. 902, pp. 61-70).

[28] Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., & Smith, N. A. (2014). A dependency parser for tweets. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, to appear (Vol. 4, No. 1.2).

[29] Owoputi, O., O’Connor, B., Dyer, C., Gimpel, K., & Schneider, N. (2012). Part-of-speech tagging for Twitter: Word clusters and other advances. School of Computer Science, Carnegie Mellon University, Tech. Rep.

[30] Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment analysis. Entropy, 17.

[31] Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903.

[32] Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 1320-1326).

[33] Gupta, M., & Zhao, P. (2012). Evaluating Event Credibility on Twitter. Proceedings of SIAM International Conference on Data Mining, 153–164.

[34] Yang, J., Arbor, A., Counts, S., Morris, M. R., Hoff, A., & Way, O. M. (2013). Microblog Credibility Perceptions: Comparing the United States and China.

[35] ISO/IEC 25012:2008(E) Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. (2008). Switzerland.

[36] Johnson, T. J., Kaye, B. K., Bichard, S. L., & Wong, W. J. (2007). Every Blog Has Its Day: Politically-interested Internet Users' Perceptions of Blog Credibility. Journal of Computer-Mediated Communication, 13(1), 100-122.

[37] Abdulla, R. A., Garrison, B., Salwen, M., Driscoll, P., & Casey, D. (2002, August). The credibility of newspapers, television news, and online news. In Education in Journalism Annual Convention, Florida USA.

[38] Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (pp. 591-600). ACM.

[39] Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012). Tweeting is believing? Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work - CSCW '12, 441–450.

[40] Schmierbach, M., & Oeldorf-Hirsch, A. (2012). A little bird told me, so I didn't believe it: Twitter, credibility, and issue perceptions. Communication Quarterly, 60(3), 317-337.

[41] Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014). Tweetcred: Real-time credibility assessment of content on twitter. In Social Informatics (pp. 228-243). Springer International Publishing.

[42] Erp, M. Van, Hage, W. R. Van, Hollink, L., Jameson, A., & Troncy, R. (2011). Detection, Representation, and Exploitation of Events in the Semantic Web. Proceedings of the 11th Interational Semantic Web Conference ISWC2011, 7 Suppl 1(October), I1.

[43] Nagy, A., & Stamberger, J. (2012, April). Crowd sentiment detection during disasters and crises. In Proceedings of the 9th International ISCRAM Conference (pp. 1-9).

[44] Vakharia, D., & Lease, M. (2013). Beyond AMT: An Analysis of Crowd Work Platforms.

[45] Aroyo, L., & Welty, C. (2013). Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. WebSci2013. ACM, 2013.

APPENDIX I: SEED WORDS

Table 16: These seed words are provided by domain experts from the social science department. Each seed word is categorized by the domain experts as event, location, actor/organization, or other.

Events	Location	Actors / organizations	Other
commercial whaling	Japan	Japan Whaling Association	harpoon cannon
whaling	shops	International Whaling Commission (IWC)	harpoon
hunting	restaurants	Institute of Cetacean Research	Markets
moratorium	North Pacific Ocean	pro- and anti-whaling countries and organizations	whale meat
quota	Southern Ocean	Nations	
	Antarctica	Scientists	
	factory ship	environmental organizations	
	factory ship Nisshin Maru	NGOs	
	security patrol vessels	United Nations International Maritime Organization	
		Antarctic Treaty System	
		Anti-whaling governments	
		Anti-whaling groups	
		Greenpeace	
		Japanese government	
		World Wildlife Fund	
		Ocean Alliance	
		Sea Shepherd Conservation Society	
		Japan Fisheries Agency	

Table 17: Mapping between "seed words" and "hashtags for tweets" and "keywords for news articles"

News keyword (non-overlapping)	News keyword (overlapping)	Twitter reference
Hunting	Japan Whaling Association	#JapanWhalingAssociation
Moratorium	International Whaling Commission	#InternationalWhalingCommission
Quota	Institute of Cetacean Research	#InstituteofCetaceanResearch whale
Japan	United Nations International Maritime Organization	#UnitedNationsInternationalMaritimeOrganization whale
shops	Antarctic Treaty System	#AntarcticTreatySystem whale
restaurants	Greenpeace	#Greenpeace whale
North Pacific Ocean	World Wildlife Fund	#WWF Whale
Southern Ocean	Ocean Alliance	#OceanAlliance whale
Antarctica	Sea Shepherd Conservation Society	#SeaShepherdConservationSociety
factory ship	Japan Fisheries Agency	#JapanFisheriesAgency
security patrol vessels	Whaling	#whaling
Japan Whaling Association	Commercial Whaling	#CommercialWhaling
Pro- and anti-whaling countries and organizations	Factory ship Nisshin Maru	#NisshinMaru

Nations Scientists environmental organizations NGOs Anti-whaling governments Anti-whaling groups Japanese government Ocean Alliance Japan Fisheries Agency harpoon cannon harpoon Markets whale meat	Commercial whaling Whaling	Commercial whaling whaling

APPENDIX II: ARTICLE MAPPING

Table 18: Mapping that normalize the fields retrieved from the NYTimes API and The Guardian Open Platform API

Mapped field	NYTimes API field	Guardian API field
title	[u'headline'][u'main']	[u'webTitle']
abstract	[u'content'] (first two sentences, nltk sentence tokenizer)	[u'body'] (first two sentences, nltk sentence tokenizer)
word_count	[u'word_count']	[u'fields'][u'wordcount']
source	[u'source']	[u'fields'][u'publication']
url	[u'web_url']	[u'webUrl']
author	[u'byline'][u'person'] (firstname, lastname) OR [u'byline'][u'organization']	[u'blocks'][u'body'][0][u'createdBy'][u'firstName'] + [u'blocks'][u'body'][0][u'createdBy'][u'lastName'] OR [u'fields'][u'byline']
keywords	[u'keywords']	[u'tags']
publication_date	[u'pub_date']	[u'webPublicationDate']
body	[u'content']	[u'body'] (extracted from webpage via [u'webUrl'])
comment_count	[u'comment_count'] (extracted from webpage using Microdata)	[u'comment_count'] (extracted from webpage using Microdata)
comments	[u'comments'] (extracted from webpage using Microdata)	[u'comments'] (extracted from webpage using Microdata)
extracted_from	http://developer.nytimes.com/docs/	http://open-platform.theguardian.com/

APPENDIX III: CROWDFLOWER TASKS

This appendix contains the three crowdsourcing task user interfaces. The following programming and scripting languages were used to create the templates:

Table 19: Programming and scripting languages used to create the Crowdflower template

Language	Full name	Reference	Description
CML	Crowdflower Markup Language	https://success.crowdflower.com/hc/en-us/articles/202817989-CML-and-Instructions-CML-CrowdFlower-Markup-Language-Overview	Provides the functionality to put Crowdflower data fields into your HTML template.
HTML5	HyperText Markup Language	http://www.w3.org/TR/html5/ https://en.wikipedia.org/wiki/HTML5	HTML5 is a core technology markup language of the Internet used for structuring and presenting content for the World Wide Web. As of October 2014 [update] this is the final and complete fifth revision of the HTML standard of the World Wide Web Consortium (W3C). The previous version, HTML 4, was standardized in 1997.
Javascript	Javascipt	http://www.w3schools.com/js/	JavaScript is the programming language of HTML and the Web. Programming makes computers do what you want them to do.
CSS3	Cascading StyleSheet	https://developer.mozilla.org/en/docs/Web/CSS/CSS3	CSS3 is the latest evolution of the Cascading Style Sheets language and aims at extending CSS2.1. It brings a lot of long-awaited novelties, like rounded corners, shadows, gradients, transitions or animations, as well as new layouts like multi-columns, flexible box or grid layouts
Bootstrap 3	Twitter Bootstrap	http://getbootstrap.com/	Bootstrap is the most popular HTML, CSS, and JS framework for developing responsive, mobile first projects on the web.

Since the interaction of the user interface is not readable from the screenshots they will be described here:

Table 20: Description of the interactions used in the Crowdflower tasks

Event	Result	Category
Clicking a "what tweet is more credible?" radio button	LEFT is MORE credible than RIGHT: - the left tweet/article header becomes green - the left tweet/article header becomes red LEFT is LESS credible than RIGHT: - the left tweet/article header becomes red - the left tweet/article header becomes green BOTH tweets are EQUALLY credible: - all tweet/article headers become grey	Usability
Clicking the submit button, without selecting any features	A pop-up message appears with the question: "Are you sure you want to continue without selecting any features?" and a yes/no choice.	Validation
Hover a feature	A description appears as a tooltip, describing the hovered feature.	Usability

The next paragraphs describe the task description and show screenshots of the user interfaces of the crowdsourcing tasks. The workers are able to leave an open comment in a text area below the features.

Tweets with author features

Task description
Below you see two tweets about "whaling" or "whale hunting". Additionally, factors related to the tweets (e.g. retweets, hashtags, author) are displayed. Answer the following questions: **STEP 1:** Compare the tweets and determine what tweet is **MORE CREDIBLE** **STEP 2:** Select the **FACTORS** that indicate the credibility of the tweet. **STEP 3:** Provide any **COMMENTS**. (optional) **CREDIBILITY** is defined as a combination of **COMPETENCE**, **TRUSTWORTHINESS**, and **GOODWILL**.

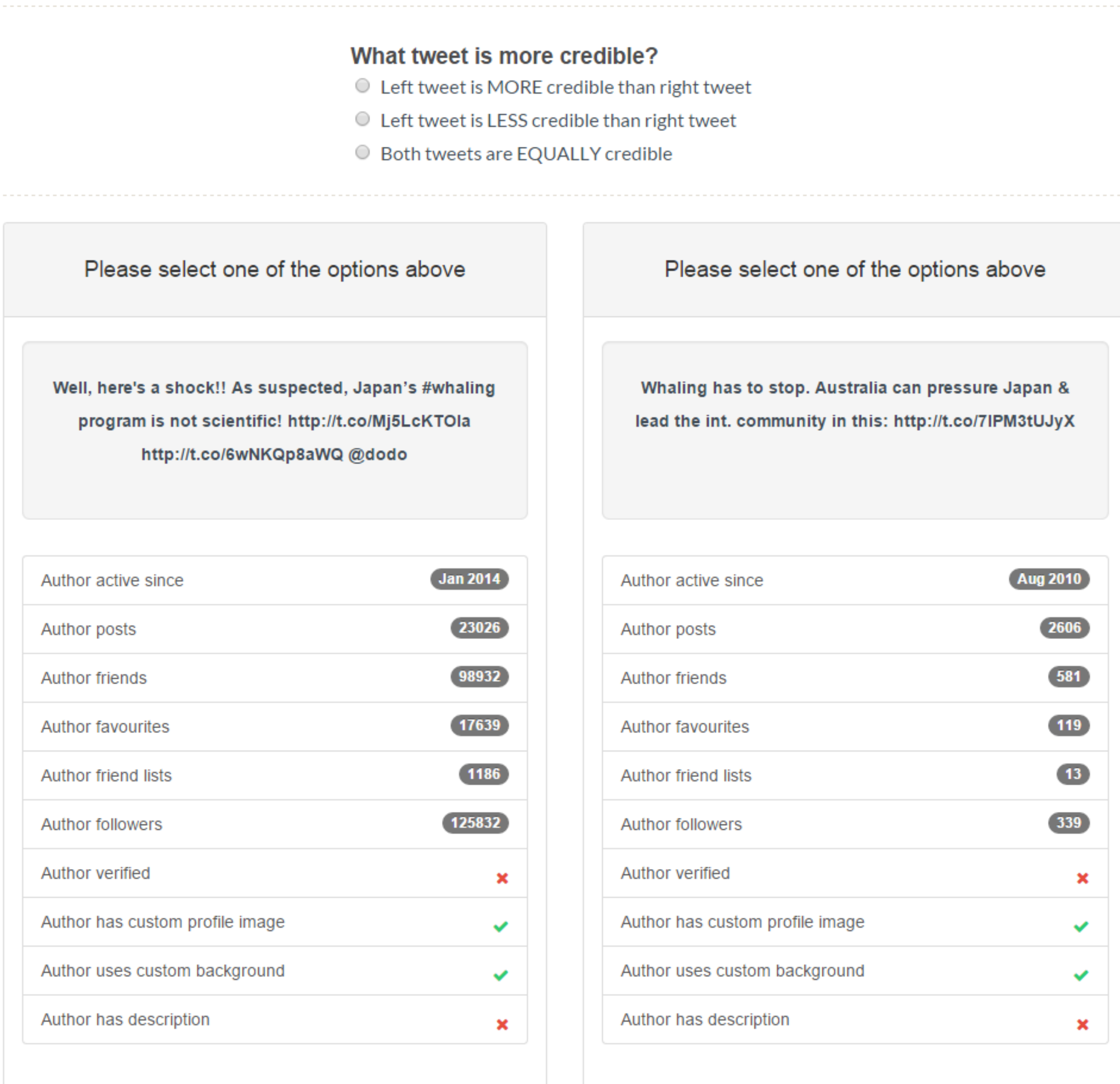

Figure 5: The Crowdflower task user interface for the tweets with author features

Tweets with other features

Task description
Below you see two tweets about "whaling" or "whale hunting". Additionally, factors related to the tweets (e.g. retweets, hashtags, author) are displayed. Answer the following questions: **STEP 1:** Compare the tweets and determine what tweet is **MORE CREDIBLE** **STEP 2:** Select the **FACTORS** that indicate the credibility of the tweet. **STEP 3:** Provide any **COMMENTS**. (optional) **CREDIBILITY** is defined as a combination of **COMPETENCE**, **TRUSTWORTHINESS**, and **GOODWILL**.

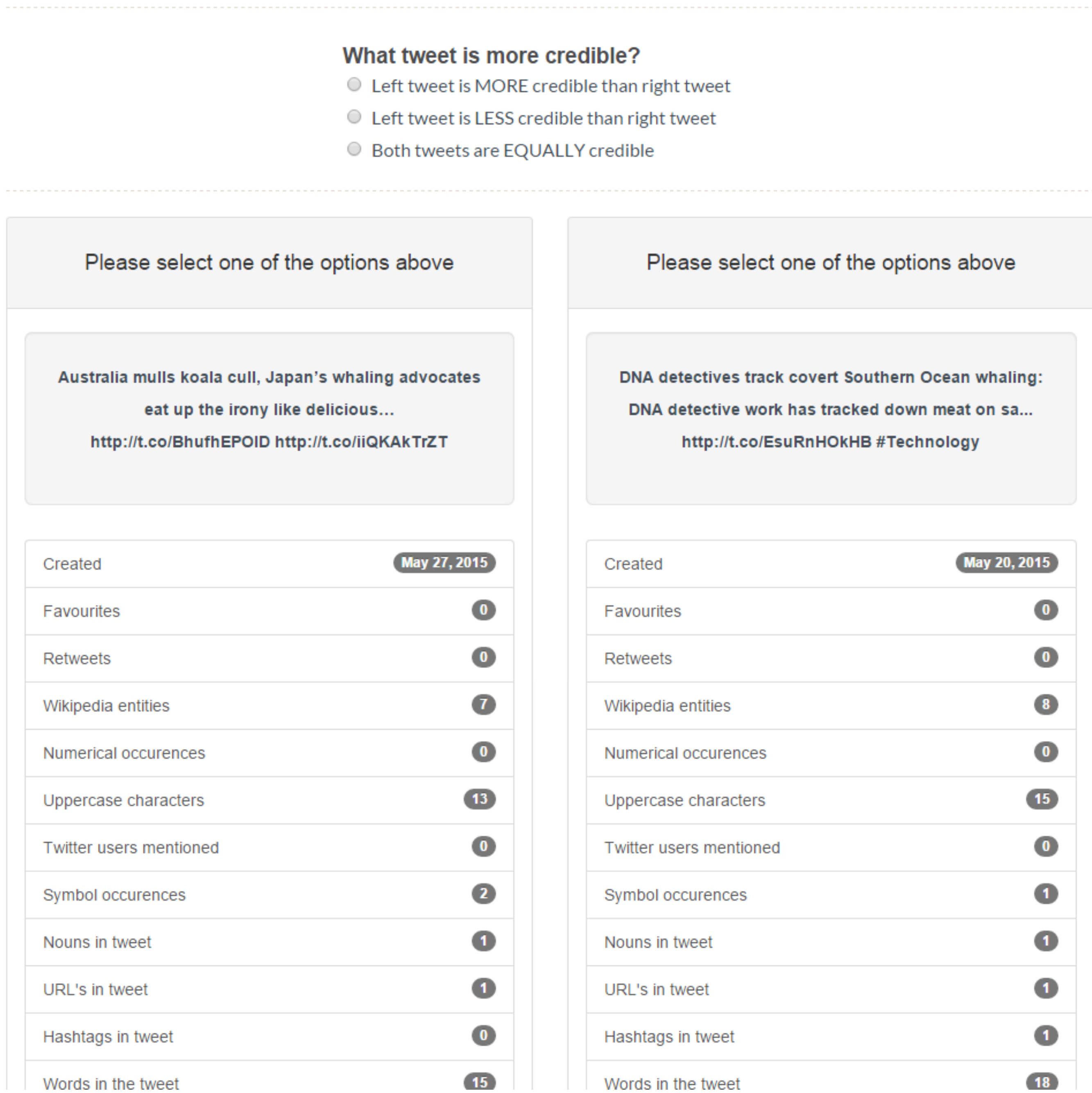

Figure 6: The Crowdflower task user interface for the tweets with other features

Articles

Task description
Below you see two article abstracts about "whaling" or "whale hunting". Additionally, factors related to the articles (e.g. comments, sentiment, word count) are displayed. Answer the following questions: **STEP 1:** Compare the articles (and features) and determine what article is **MORE CREDIBLE** **STEP 2:** Select the **FACTORS** that indicate the credibility of the article. **STEP 3:** Provide any **COMMENTS**. (optional) **CREDIBILITY** is defined as a combination of **COMPETENCE**, **TRUSTWORTHINESS**, and **GOODWILL**.

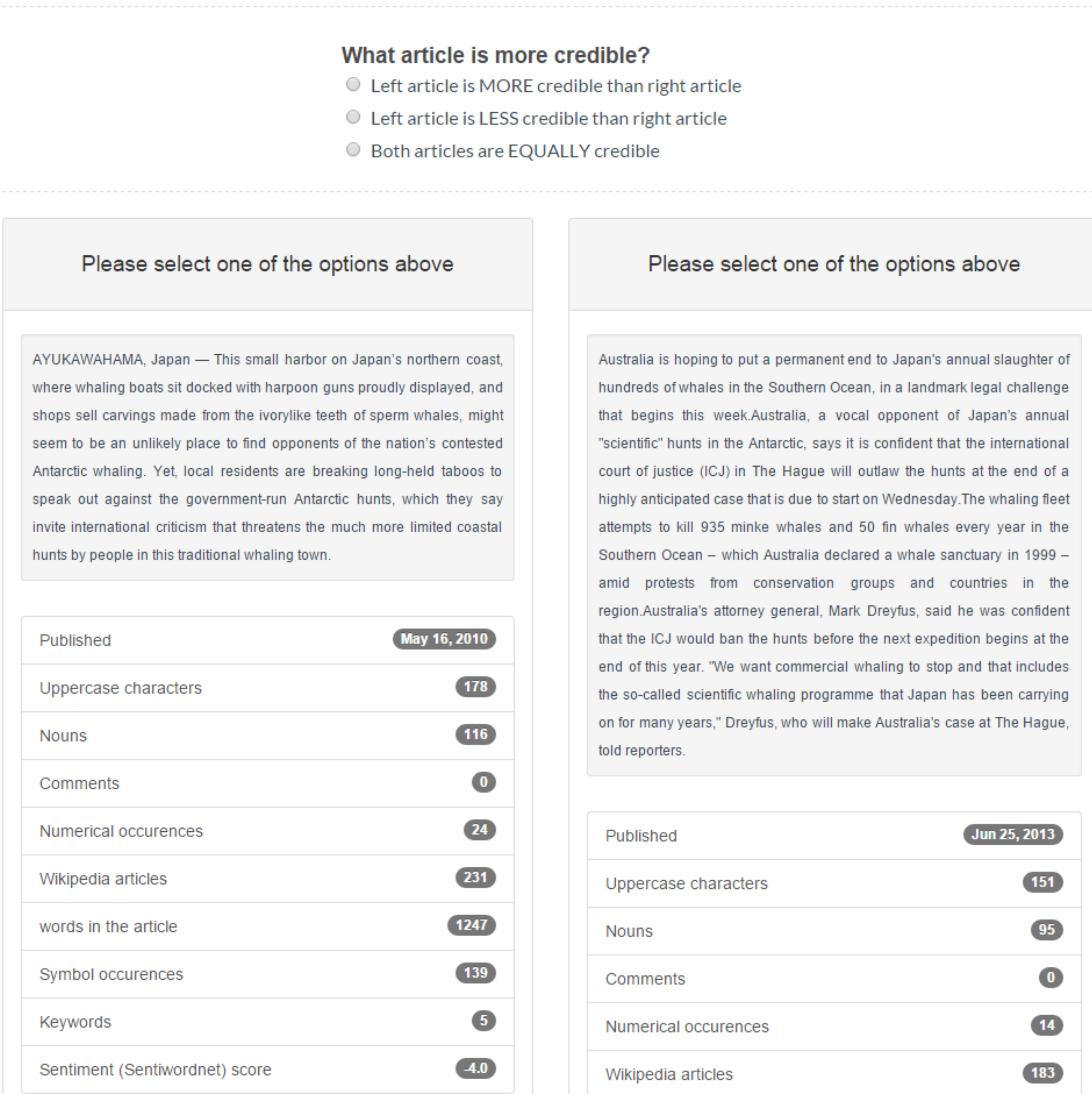

Figure 7: The Crowdflower task user interface for the articles

APPENDIX IV: COMMENTS

During the crowdsourcing task the work was asked to optionally provide comments. The tables below show the provided comments, together with the credibility factors extracted from the comments. The grey colour of a credibility factor cell means that we were not able to extract any credibility factors from the article or tweet.

Comment (Articles)	Credibility factor
Again, both articles appear trustworthy; and again, I am unable to select the phrases which lead me to this conclusion, so here they are below: First article The International Court of Justice has ordered a temporary halt to Japan's annual slaughter of whales in the Southern Ocean The UN court's decision Second article member governments of the International Whaling Commission (IWC), the whaling regulatory body commercial whaling would be barred	
As before the tone of the article lends credibility for me.	Tone
As far as I can see, both articles are reporting the decision of the Court in a neutral manner and appear to be totally trustworthy as far as the facts are concerned. I have tried to select credibility factors by clicking on the phrases, but this doesn't work; so I will list them below. First article The decision to ban Japanâ€™s annual whaling drive off Antarctica United Nationsâ€™ highest court International Court of Justice in The Hague Second article decision by the International court of justice ordering a halt to Japanâ€™s whaling program in the Southern ocean The UN court	Objectivity
author spoke in complete sentences and gave more detail	Completeness
Both are factual, non opinionated articles	Factuality
Both articles are perfectly credible. I base this conclusion largely on the following: First article International Court of Justice (ICJ) ordered a halt Tokyo said it would abide by the decision Second article international court banned its controversial "scientific" whale hunts in the Southern Ocean a vow by Japan's prime minister, Shinzo Abe, to make efforts to restart commercial whaling despite a ban by the international court of justice	
Brevity was a factor here	Brevity
clear explanation	Clarity
Different views/reports on a related event	Multiple viewpoints
Fewer wikipedia articles suggests more primary source material perhaps	Fewer Wikipedia entities
Godd stories both	
I don't really get the stats and what they have to do with cerdibility, they may or may not be true, but char counts means what? your score means what? I't usless infomration	
I wonder what the comments are in the second article? I get the impression that they are rejecting the comments made in the article and it is for this reason I have given it less credibility.	Content of the comments
i would like to think everyday consumers would be against whale hunting	
it gives more info	Completeness
looks like more important	Importance
MORE UP TO DATE MORE INFO	Novelty, completeness
much more complete article	Completeness
non of the stats are relevant, what is the connection between word count and credibility supposed to be?	
non of the stats are relevant, what is the connection between word count and credibility supposed to be? It very nice to count all that, but a well written lie is still false, regardless of any of those	
non of the stats are relevant, what is the connection between word count and credibility supposed to be?	

It very nice to count all that, but a well written lie is still false, regardless of any of those numbers	
Quotes were specified.	Quotes in article
summarization exact	
The article on the right is credible and trustworthy, a decision which I base largely on the following phrases: An anti-whaling activist was detained by the crew of a Japanese whaling ship Pete Bethune, jumped aboard the Shonan Maru No 2 this morning from a jetski driven by Larry Routledge, a British campaigner, Sea Shepherd said. Hang on - you aren't really expecting me to decide on the credibility of these articles by the weird statistics underneath them, are you? Surely one can only base such decisions on the text itself, rather than how many pieces of punctuation are used or totally erroneous things like "numerical occurrence"? No, you can't be expecting that!!	
The fact that in the second article there is no mention of how many whales Japan will be targetting means that this article is less credible to me.	Numerical occurrences
The fact that it is a Conservation group that is claiming that the meat is unfit for human consumption, suggests to me that there is less trustworthiness and credibility because of what the group represents.	Objectivity
The left article is more credible, since it cites sources - something which the right article totally fails to do. Examples: the government said The Agriculture Ministry	References
the right article goes into MUCH more detail about the subject. the left article seems more like a summary than an actual article.	Level of detail
the right article seems better educated. the left article definitely doesn't have 546 words like it says.	
the right one seems to go into more detail about actual people involved with whale hunting	Level of detail
The tone appears to be reporting rather than having an emotional bias from what is given.	Tone
This one was easy as #2 relates to incidents from @ 150 years ago.	Novelty
Why no comments on either?	

Comments (Tweets with author features)	**Credibility factor**
#1 INFORMATION AS OPPOSED TO A PLEA	Factuality
#1 just a ramble	
#2 MORE TO OFFER	Completeness
almost human like stats, maybe we have a genius person here?	Human-like stats
AUTHOR FOLLOWERS IS MORE	Author followers count
AUTHOR FRIENDS IS MORE	Author friends count
AUTHOR FRIENDS IS MORE	Author friends count
AUTHOR FRIENDS IS MORE	Author friends count
author good	
Author on left is a US SEnator!	Authority
Basically the same tweet	
been around longer	
Bot stats, right mite be a normal user, MAYBE	Human-like stats
both articles are credible and informative	
Both articles are reasonable	
both equal, yes, not NOT creditable at all ...post count to high	Post count too high
both equally, and both meaningless ...at best, it is good that the authors is NOT verivied, so it's not a payed writer	
Both have solid info	
both make money with posts, credibility = 0	Objectivity

Close call here	
credibilty = 2/5, stats are possible, but high for a genuine user	Human-like stats
equally creditable, where credibility = 0.0. Pro writer, making money	Objectivity
facts not verifie	Evidence
have not verified facts stated	Evidence
I think the tweet has better conformation and the information the author gives validity	Evidence
in just over one year activity, way over 20 000 posts ...= pro = payed = $$ => credibility = 0	Human-like stats
left tweet is not about whaling	
Left tweet makes a serious point and does it clearly, the right tweet is like an advert	Clarity, style of the tweet
link makes it more credible	Contains hyperlink
Looks's like the same author	
No clue what the right tweet is actually saying!	Clarity
not verified, and low stats, it MITE be a normal human	Human-like stats
nothing obtuse or controversial in either article	Controversy
presented as reportage.	Style of the tweet
Provides evidence	Evidence
Right tweet is a sea of hieroglyphics, left tweet just states the facts	Symbolic occurrences
silly icons	Symbolic occurrences
THE AUTHOR FRIENDS IS MORE	Author friends count
the author on left uses 4 instead of four	Written numbers
the claim on the right side is too difficult to complile	
the comment on the left is useless	
THe comment on the Left shows the shortfalls of social communication	
the jerk on the left is just an agitator!	Sentiment
The left tweet is to the point, the right tweet is just a teaser for the conference	Focus
The right tweet has nothing to do with Japan whaling	
The right tweet is not about whaling	
The second tweet hardly makes sense, so clearly lacks credibility	
The tweets are identical, just retweets, so equally credible	
there seems to be a certain sence you try to assume that if the author is verified ....he's not a bot, payed writer, etc. there is no value in 80-% of your stats. YOu only know if they are low-missing, it is NOT a pro. If the stats are there, you know nothing.	
This was a tough one; in the end I went with the second one because although they have less followers, they have put significantly more posts into their account, and have a higher number of friends and customization on their account.	Combination of Profile customisation, author followers count, post count, firneds count

Comment (tweet with other features)	**Credibility factor**
again, to me the stats provided are useless toward credibility	
almost no content ... so credibility of what? compared to what?	
although 1st more personal, both are credible	Objectivity
Any tweet which uses the word "fail" in that way, when talking about something important, immediately loses credibility points!	Sentiment

Both have too many hashtags	Too many hashtags
Can't read the other one, i do not know japanese	
comment on right simply states a fact	Factuality
I don't get this, credibility of what, compared to what? a poem? If you don't provide content in form of articles or posts, what are we suppsoed to judge here?	
I stil fail to see how char count should add to credibility	
Is that really Michelle Obama	
left is emotional, right is state of fact	Sentiment
Left tweet is not in English	
left tweet is ridiculous	
None of the factors are important, the tone of the right tweet is micky taking which loses it a lot of credibility	Sentiment
not even sure if left is meant to be true!	
One is not in English	
right tweet have sustain information	Sustainability
right tweet looks more serious	Sentiment
Right tweet loses it's credibility by using mockery!	Sentiment
snarkiness in the one on the right	Sarcasm
support information	
The left tweet uses better English and is quoting experts. The right is a somewhat childish plea to Barack Obama, which is unlikely to achieve anything. None of the factors really manage to identify any differences in credibility.	Spelling/grammar
The left tweet's in Japanese, which makes it difficult! In translation it seems they're looking to make excuses, which lacks credibility.	
The tweet on right is from Japanese whalers	Objectivity
they are both equal, and have no creditiblity at all. THe data provided are not jelpfull in any way. What has symbol count to do with credibility?	

APPENDIX V: CORRELATION PLOTS

The correlation for every credibility feature is represented in a set of scatter plots, with the "times a worker indicated a feature as credible, in respect to the tweet or article", on the horizontal axis, and the "times a tweet or article is indicated by the worker as more credible" on the vertical axis. The "r" below the plot corresponds to the Pearson correlation coefficient and consists of a number between 1 (high positive correlation) and -1 (high negative correlation), which means values close to zero correspond to a low correlation.

Figure 8 describes the correlation of ten features drawn from the results of the "tweets with author features" crowdsourcing task. Each scatter plot contains 99 tweets, which are represented as a small circle. The small circles may overlap, so one circle may actually represent multiple circles.

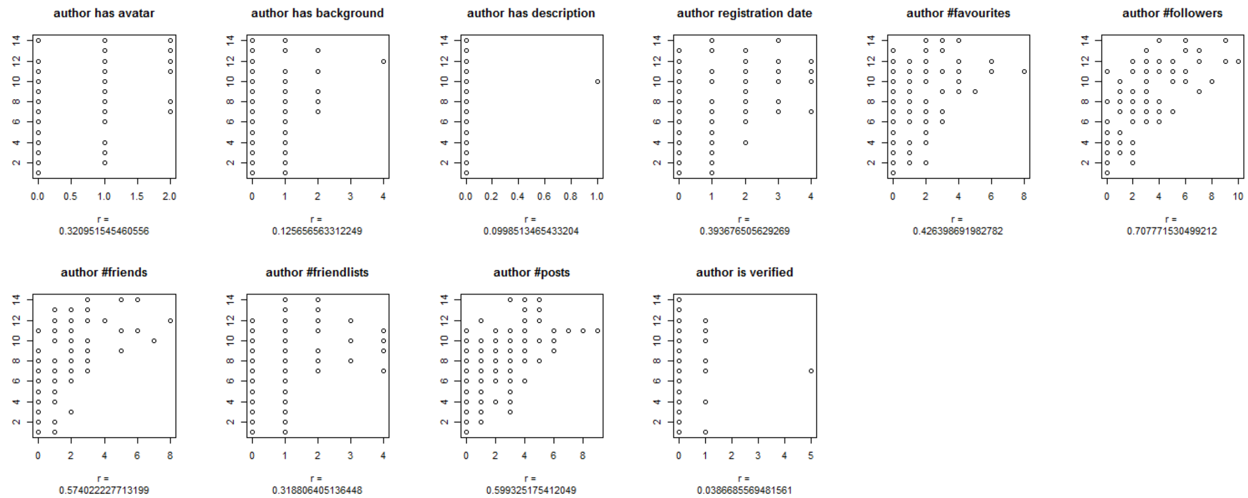

Figure 8: Correlation scatter plots representing "#tweets marked as more credible" versus "#tweet author features marked as an indication of the credibility"

Figure 9 describes the correlation of sixteen features drawn from the results of the "tweets with other features" crowdsourcing task. Each scatter plot contains 99 tweets, which are represented as a small circle. The small circles may overlap, so one circle may actually represent multiple circles.

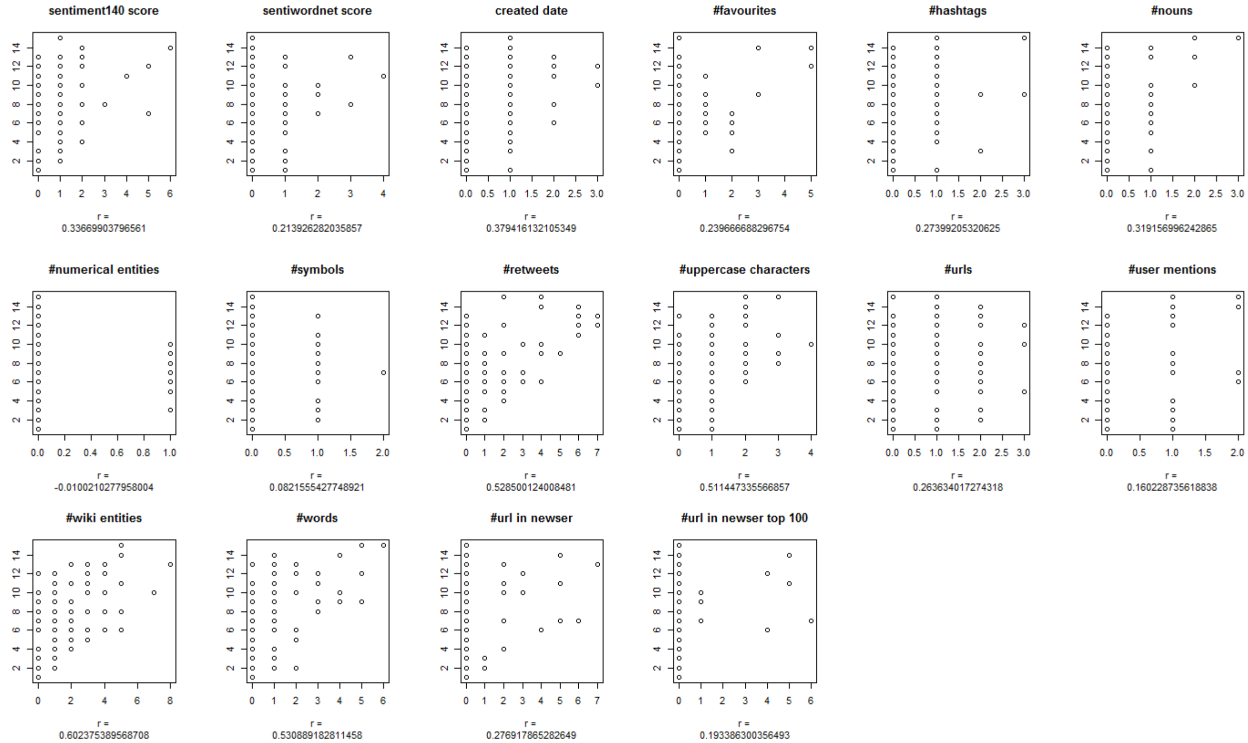

Figure 9: Correlation scatter plots representing "#tweets marked as more credible" versus "#tweet other features marked as an indication of the credibility”

Figure 10 describes the correlation of ten features drawn from the results of the “articles” crowdsourcing task. Each scatter plot contains 100 articles, which are represented as a small circle. The small circles may overlap, so one circle may actually represent multiple circles.

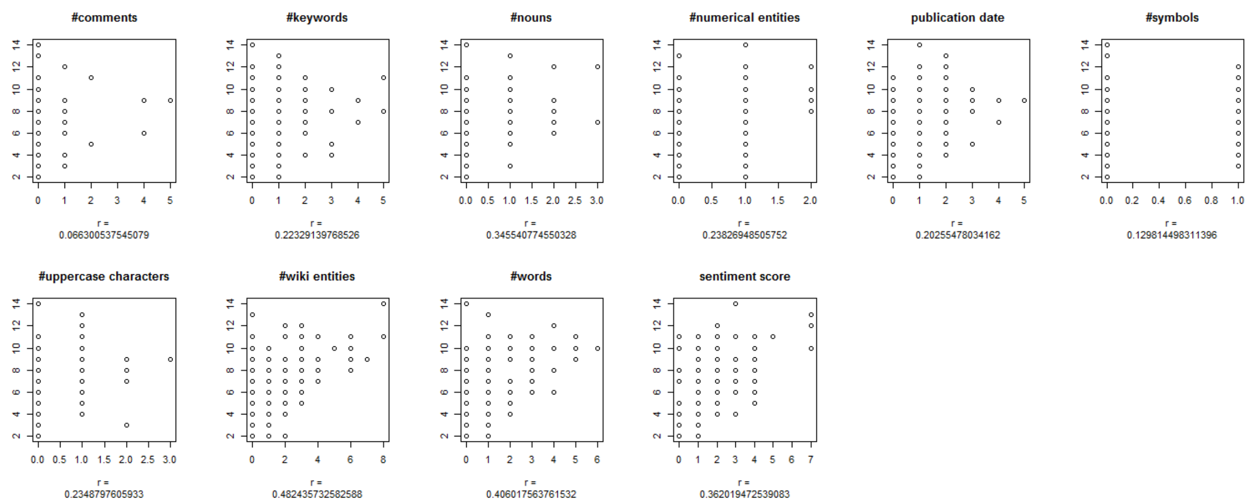

Figure 10: Correlation scatter plots representing "#articles marked as more credible" versus "#article features marked as an indication of the credibility”

APPENDIX VI: FEATURE VALUE DISTRIBUTIONS

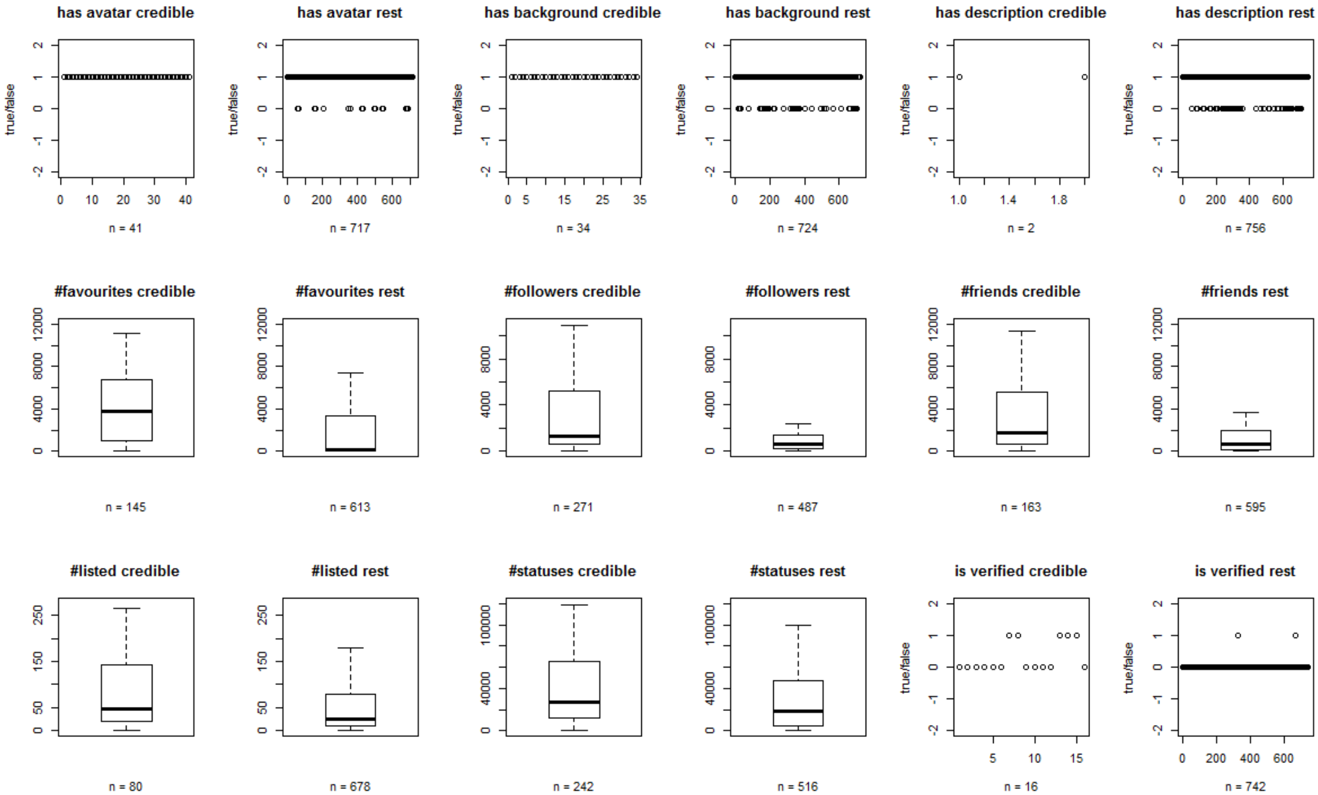

Figure 11: scatter plots and boxplots of the values of the author features of a tweet

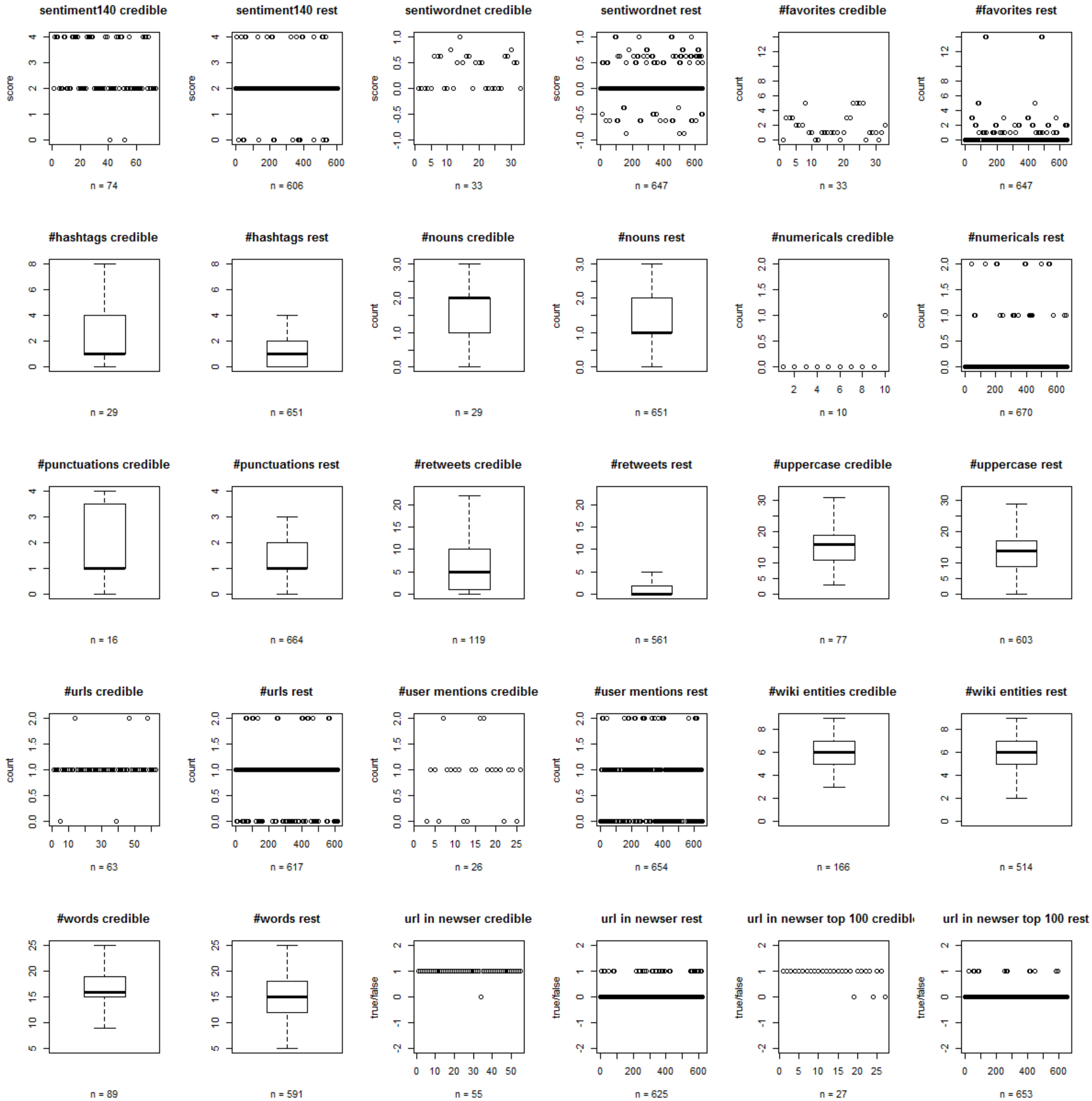

Figure 12: scatter plots and boxplots of the values of the other features of a tweet

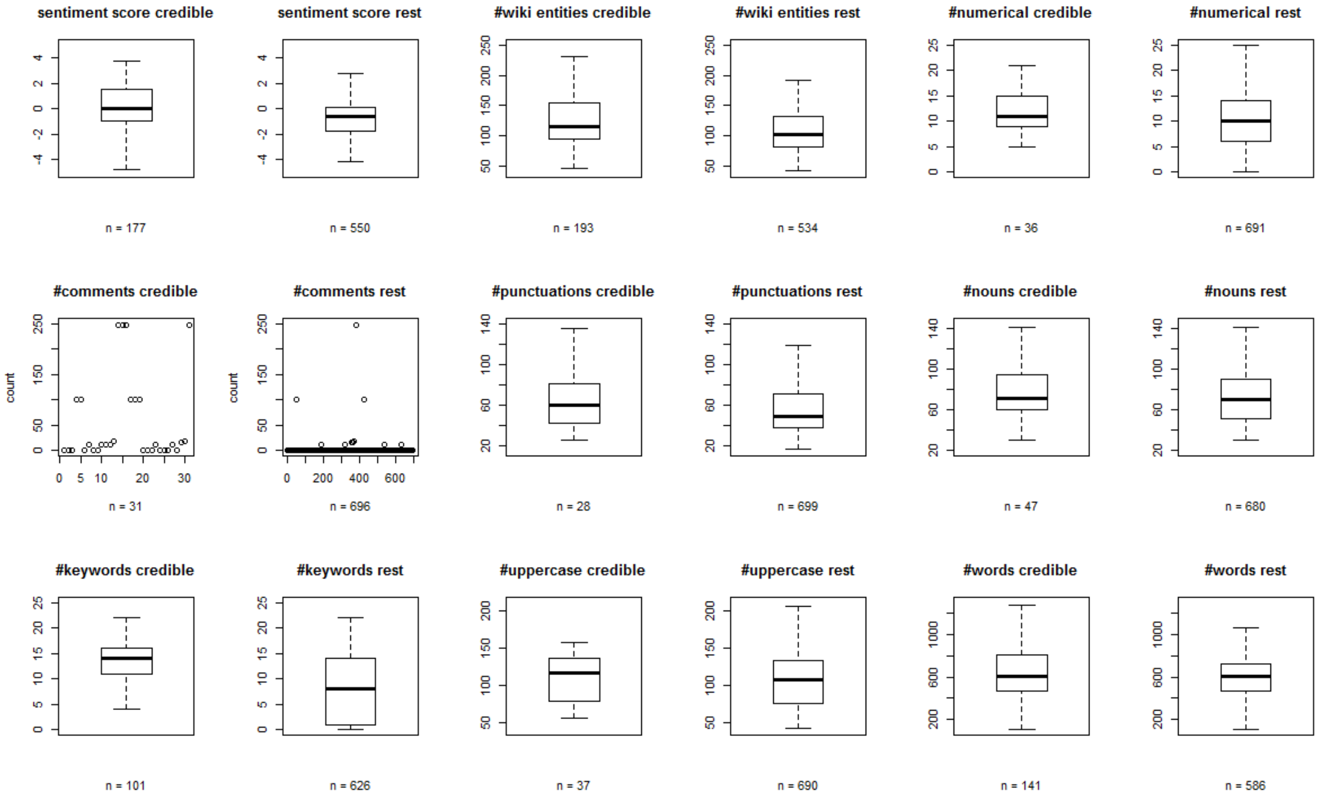

Figure 13: scatter plots and boxplots of the values of the features of an article