A Digital Nudge to Fight Confirmation Bias and the Spread of Misinformation

Quentin Meeus, Calum Thornhill

January 29, 2019

Contents

Introduction

Fake news exploits individual's subjective reading of news in order to spread. Sometimes differentiation between real and fake information is apparent. However, it is often the case that a message is written to evoke certain emotions and opinions by taking partially quasi-true base stories and injecting false statements such that the information appears to be realistic.

It was shown that in the months preceding the 2016 American presidential election, organisations from both Russia and Iran ran organised efforts to create such stories and propagate their spread on Twitter and Facebook. This presents an arguments for the importance of raising internet users' awareness of such practices. Key to this is providing users with means to understand whether information should be trusted or not.

A solution put forward by social networks proposes to rely on users to identify suspicious articles shared on their platforms. The flagged articles are subsequently fact checked by third-party volunteers. Then, when a user comes across such article, he is given the chance to read an alternative article that has been deemed trusted. In this work we propose a low-cost method for fighting the spread of fake news without having to rely on third parties or users.

The method that we are proposing presents the user with a selection of articles from a range of reputable news sources providing alternative opinions from the whole political spectrum. This strategy is a form of digital nudge, in which the user is presented with an original text together with articles proposing alternatives views placed at eye level.

Our main objective is that such a tool will educate people about sharing and believing information accessed online which in turn can decrease the spread of fake news. We also hope to increase awareness regarding the different ways an information can be presented and sometimes manipulated online.

In Section 1, we briefly discuss the mechanisms of misinformation spreading online and how social networks are the perfect platforms to accelerate this process. In Section 2, we take a closer look at the related research in the field of fake news and nudge design. In Section 3, we go more into the details of the nudge strategies considered, following what was learned in Section 2. We briefly describe our approach in the big lines in Section 4 and give a first high level description of our tool and provided use cases and examples in Section 5. Technical design and the inner workings are covered in Section 6. We end up with a description of the experiments that we have conducted in Section 7 and conclude with the future research directions.

1 Spreading Mechanisms of Online Misinformation

In this section, we discuss how social networks increase the spread of biased news and misinformation. We discuss confirmation bias, echo chambers and other factors that subconsciously influence a person's opinion. We show how these processes interact together to form a vicious circle that favour the rise of untrustworthy sources.

1.1 Confirmation Bias

Oftentimes, when we think we know something, we are satisfied by an explanation that confirms our belief without necessarily considering all possible other explanations. This is confirmation bias in action. Nickerson [10] defined it as the tendency of people to both seek and interpret evidence that support an already held belief. In other words, people tend to believe an information that confirms their belief with no or few regards to its veracity. As the saying goes: *"It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so."*[1].

1.2 Echo Chambers

An echo chamber is a situation in which the only things that an individual can hear is an echo of something that has just been said [3]. Social networks such as Twitter and Facebook are environments that favor largely the creation of such chambers [4]. People tend to mix with the ones who think like them and follow news sources they favour. In doing so, people see one side of the argument and are not offered the opportunity to a whole vision of the opinion landscape.

Consider a user with a hard line political belief on either side of the spectrum. They may follow only people and news sources who share that belief. It would likely be the case that upon publishing a tweet about a new policy or event, they would see similar tweets from their friends within and receive feedback that favours their own opinion. The echo chamber around the user acts as a wall between him and a holistic opinion. The 2016 American presidential election illustrates very well this phenomenon. Donald Trump's victory came as a surprise to many people worldwide. One explanation of this surprise is that web users were enclosed in echo chambers.

Research and having a critical approach to information shared online can protect a user against biased views but very few protections exist against the creation of echo chambers. People can learn to identify them but to avoid it completely means to make sure that all opinions are represented within one's social circle. Even if someone manages to accomplish that, the next section will show that it is no guarantee for unbiased information.

1.3 Recommender Systems

Social networks extensively use algorithms for proposing content. The reason is simple: the amount of content available online would completely overflow us if we were presented with everything everytime. Also, social networks want to improve the user experience by displaying content that they will appreciate. Unfortunately, this is only worsening the problem at hand because it implies to group users into clusters of preferences and providing them with filtered content.

Such algorithms are called recommender systems and rely mostly on artificial intelligence to decide which content is best for a particular user. They are based on two common techniques: content-based

[1] Although this quote is often attributed to Mark Twain, there are serious doubts regarding the veracity of this affirmation [6].

filtering and collaborative filtering, although often hybrid models are built by mixing both techniques. In the former case, website engines build a profile of what the user might like based on his previous interactions with the website, for example what articles he clicked on or what content he liked. The second technique builds clusters of users that have been known to like and share the same opinions. The algorithm then shows content liked by other users from the same cluster. When in presence of new content, the algorithm makes inferences based on related items and decide which clusters are likely to notice it and which clusters will ignore it.

Both of those models further increase the confirmation bias and feed echo chambers. Indeed, this is a key part of the functionality of platforms, users are provided with content that they will like by restricting articles that may not encourage further interaction.

An interesting parallel to make here is with the Bayes' rule, which states that when confronted with evidence, we update a prior belief with the additional knowledge provided by this evidence. Most algorithms are somewhat derived from this rule, and recommender systems are no exception. This suggest that algorithms may suffer from confirmation bias as well, and that this is not something specific to human-kind.

1.4 The Infinite Loop

We see that these phenomena complement and feed from each other in a vicious circle. Indeed, echo chambers arise both from user's subconscious choice of surrounding himself with like-minded people and from content that is tuned to please the users' liking by recommender systems. Having access to only one side of the truth further increases the confirmation bias that what they believe is right. Finally, the fact that users respond to articles that they like closes the loop by feeding the recommender algorithms that provide them with content.

1.5 Research Question

We have mentioned in the introduction that nudges were a clever trick to influence people into making the right decision. We will define it more precisely in Section 2 and following, as this concept will have a central role in our strategy to fight fake news. However, we can already mention the question that will drive the analysis conducted in this paper:

"How can digital nudges be used to fight the spread of misinformation and online bias in social network news feeds?"

2 Related Work

2.1 The Science of Fake News

As Lazer et al. [5] sets forward, there is a scientific approach required in finding a solution to fake news in social media. Homogeneous social networks allow polarisation and closure to new information. Consequently, echo chambers would form because of the personalisation of political information. An additional reason for their formation is linked to both human behaviour and the technical foundations of the user experience.

Despite the intellectual high ground taken by fact checkers such as PolitiFact and Snopes, they do not solve the issue that is the tendency of individuals no to question efficacy of sources unless their own values or beliefs are infringed. This suggests that it is unlikely that a user would actively engage in the fact checking process and use the services provided by these fact checkers. Instead, the authors argue that it is the responsibility of platforms to include signals as to the quality of a source or article within their algorithm, for example the prioritisation of reputable sources in the news feed. However, this does not solve misinformation nor ensures that conflicting views are available to the user. A solution still remains to be found to break these mechanisms. The ideal solution should be included within the user experience and require minimal user effort.

2.2 Digital Nudges

The day-to-day definition nudge as defined by Thaler [13] is "*to push mildly or poke gently in the ribs, especially with the elbow*" or applied to an economical context, "*self-consciously attempting to move people in directions that will make their lives better*". In a digital world, the definition is no different: the idea is to influence someone's behaviour into acting in such a way that will improve his or her user experience.

Lazer et al. [5] cites nudges as a reasonable solution to the problem exposed in Section 2.1. If the reading of news on social media platforms is re-framed as a choice of choosing to believe without validation, it is possible to define an architecture around this choice. Thus, it is possible to alter this architecture through implementation of a nudge.

Several researchers have put effort into understanding the impact of nudges in social media. Acquisti et al. [1] and Wang et al. [14] have both considered nudges to encourage user awareness of privacy and the impact of posts on platforms. Acquisti et al. [1] discusses nudging by means of information, presentation, defaults, incentives, reversibility, and timing. Two strategies for nudges are presented: nudging with information and presentation.

Nudging with information involves providing information to raise awareness. For example, in the context of fake news, this may include giving a label or signal about the reputability or the political bias of a source.

Nudging with presentation involves the framing and structure of a choice. In the context of reading news in social media, an example could be the placement of an article in relation to the story from across the political spectrum.

Mirsch et al. [8] define a development process for digital nudges that is easily applicable to the choices surrounding reading news on social media platforms. Figure 1 is taken from the paper of Mirsch et al. [8].

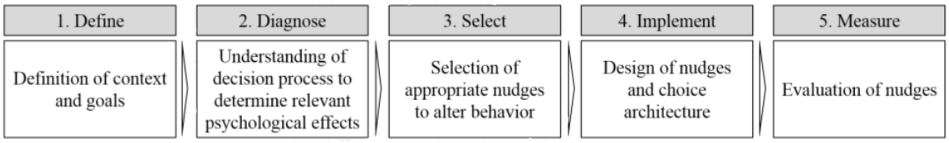

1. Define
Definition of context and goals
2. Diagnose
Understanding of decision process to determine relevant psychological effects
3. Select
Selection of appropriate nudges to alter behavior
4. Implement
Design of nudges and choice architecture
5. Measure
Evaluation of nudges

Figure 1: Nudging Development Process [8]

3 Nudges

The objective that we are trying to achieve is to induce a desired behaviour for a user where he or she is aware of the potential political bias hidden in an article and is more likely to gather information in order to assess correctly the veracity of an information. We focus on tweets posted on Twitter and discuss two approaches: nudging by presentation and nudging by information. In the first case, the user is directly presented with information that might affect his judgement. In the latter case, a visual cue is displayed that gives the user an idea on the veracity of an information. Both approaches follow the development process displayed in Figure 1. We briefly summarise the process in the context of the present paper and the proposed solution below.

- **Define**: The context is defined as the news feed of a social media platform. In the environment, only one-sided opinions are visible in the personalised sources chosen by a user. The goal is to ensure that at all times, without restricting a user from viewing the original content, the user is encouraged towards viewing a balanced representation of opinions on a subject.
- **Diagnose**: In understanding the decision process, a number of questions can be identified that would ideally be asked by any reader of news such that a reasonable investigation of reputability and veracity of source or story is made. That is, given the set of questions that a professional fact checker would ask, is there a change in the choice architecture that would encourage a non-fact checker to ask similar questions.
- **Select**: For the scope of this work, a nudge displaying a balanced view of news will be implemented. However, nudges aiming to immediately signal the veracity of a news source or article will also be tested for effectiveness. More precisely, in line with the methods of nudging of Acquisti et al. [1], a nudge by presentation will be implemented, and compared to a nudge by presentation.
- **Implement**: Implementation of the nudge will be detailed in a later section. However, in the scope of this work, the nudge will not be implanted into the user experience. Instead an emulated environment is implemented for testing that the natural language processing required for the nudge performs sufficiently.
- **Measure**: Evaluation of the nudge will rely on a user survey. The survey will ask a user to rate an article based on perceived levels of truth and reputability, in presence and absence of nudges. This is addressed in Section 7.

In the rest of the section, we go more into the details of both approaches.

3.1 Nudging by Presentation

The primary aim of the nudge is to present an unbiased view of a subject, without necessarily forcing a user to embrace it. The secondary aim, of equal importance to the first, is to ensure that the sources presented are of a level of reputability, where even if headlines are sensationalised, fundamentally the underlying article will not be entirely fictitious or propaganda.

Select the appropriate nudge The objective is to present a user with the necessary information to judge of the veracity of an information. For this reason, a selection of relevant articles are proposed in a eye-pleasing display.

Implementation Natural language processing is used to extract meaning from a tweet and the News-API is queried. These results are sorted by relevancy and presented to the user. This is described in high-level perspective in Section 4 discussed in more details in Section 6.

Presenting the nudge The alternative news sources are displayed directly below the original content. This does not restrict the user from reading the original content but achieves the purpose of placing the alternative view at eye level. This is shown in Figure 2.

Selecting News Sources A key aspect to discuss is how the sources are selected. The perceived political affiliation of news sources is identified through the reports from Pew Research [9] and YouGov [12]. They have performed a impressive work but gathering information about American and British news providers, including survey information about public opinion and perceived political bias. A set of trustworthy news sources has been extracted from their findings and is used in the tool. They are listed in Table 1 for reference.

Left	Centre	Right
The Guardian Independent MSNBC Politico	Reuters The Financial Times BBC News The Wall Street Journal CNN Bloomberg	The Telegraph The Daily Mail Fox News

Table 1: A selection of trustworthy news provider

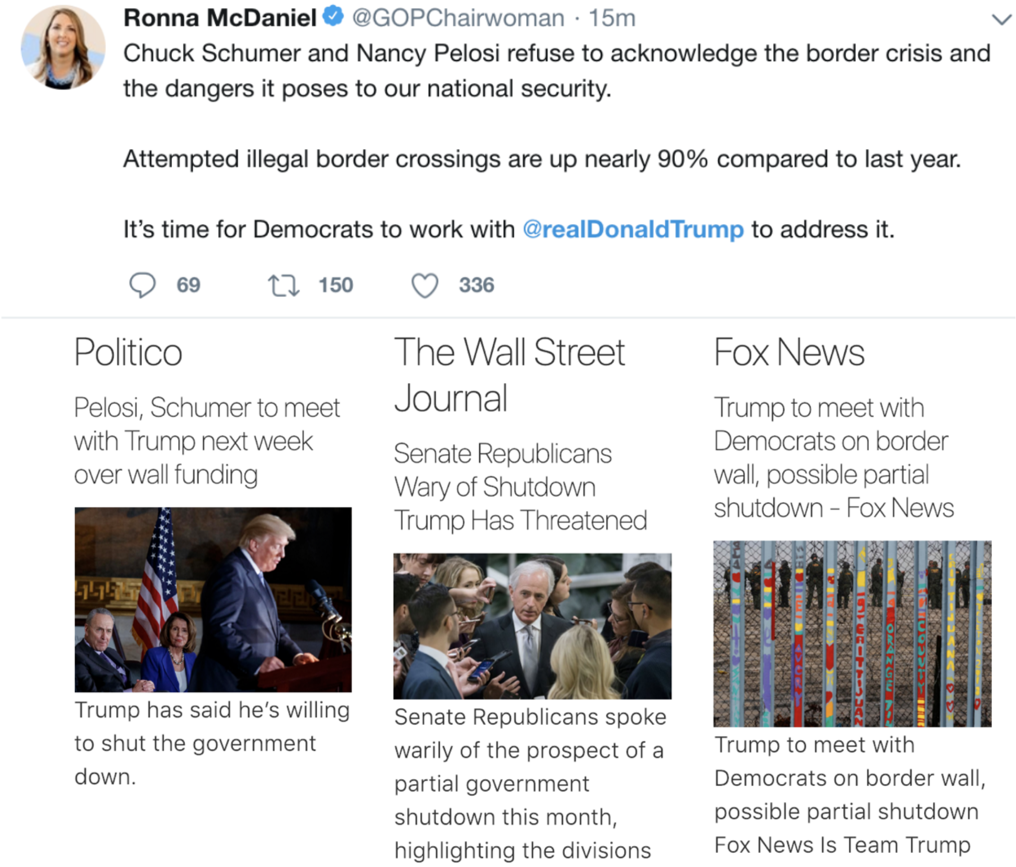
3.2 Nudging by Information

The second approach aims to provide information to raise awareness. For example, in the context of Fake News, this may include giving a label or signal about the reputability of a source or its political bias. Twitter has implemented this to some extent by showing some accounts to be 'verified'. In the presence of such an account, users see a blue and white tick. As will be explained later in Section 7, we test how an information is appreciated in presence of this visual cue when it is biased or untrustworthy. An alternative but similar approach is to include a visual cue that indicates that an account has **not** been verified.

Select the appropriate nudge In the present work, we use the existing Twitter flag for unverified accounts. It consists of a small white cross surrounded by a red background. It does not necessarily suggest bias or lack of reputability but it is the antithesis of the current nudge. We have chosen to keep this format because it is known by the users.

Implementation There is no implementation required for this test within the scope of this work. However, it will be one of the topics addressed in the experiments in Section 7.

Presenting the nudge The visual cue is placed directly adjacent to a users profile name, meaning that is likely it it's presence is acknowledged upon initial reading of a tweet.



Ronna McDaniel @GOPChairwoman · 15m

Chuck Schumer and Nancy Pelosi refuse to acknowledge the border crisis and the dangers it poses to our national security.

Attempted illegal border crossings are up nearly 90% compared to last year.

It's time for Democrats to work with @realDonaldTrump to address it.

Politico

Pelosi, Schumer to meet with Trump next week over wall funding

Trump has said he's willing to shut the government down.

The Wall Street Journal

Senate Republicans Wary of Shutdown Trump Has Threatened

Senate Republicans spoke warily of the prospect of a partial government shutdown this month, highlighting the divisions

Fox News

Trump to meet with Democrats on border wall, possible partial shutdown - Fox News

Trump to meet with Democrats on border wall, possible partial shutdown Fox News Is Team Trump

Figure 2: Placement of the nudge relative to original text

4 BalancedView: A Weapon against Online Bias and Misinformation

In the context of fighting confirmation bias and fake news in the Twitter news-feed, several mandates can be imagined. Removing all suspicious posts for example, would be an example. Another would be to not allow users to post political views that are judged too extreme. This second example reduces the platforms usability. Instead, a solution must be more subtle and not restrict a user from posting or reading any particular post. In the present section, we develop our approach and what we are proposing.

4.1 Overview

We propose a solution, nicknamed *BalancedView*[2], that aims to encourage users to consider the wider view surrounding an information. In a first proposal, we will focus on tweets from the well-known social platform Twitter[3]. We aim to implement a tool that efficiently presents a full view on articles from relevant sources presenting opinions from everywhere in the political spectrum. Practically, a user would input a tweet and be shown articles from trustworthy sources reporting on the same topic but with different opinions.

By doing so, a user is given the opportunity to forge their own opinion by reading multiple presentation an information. They can then make an informed decision on whether to believe an article based on presented alternatives. The proposed nudge is equivalent to placing the healthier bananas at eye level alongside an unhealthier option. The aim of the nudge is to ensure that a reader of a post is not restricted from reading the original content and is instead given a balanced view of the information based on sound journalism. Rather than restrict content and usability, we propose to place a balanced and reputable selection of news sources at eye level to a news.

An initial proposal is that a user will input a tweet to the tool that extracts the relevant text to structure a query to an API of all news sources. The user is then presented with multiple sources that a present a varied view of the same subject. The sources can be presented with their political affiliation, their perceived veracity, and the agreement with the entered statement. Further work would then aim to embed this system into the user experience within Twitter.

4.2 High-level Description

When a user inputs a tweet, the system first extracts and summarises into relevant keywords the information contained in the text by using the TextRank algorithm [7]. With the keywords, the system builds a query to search for articles using the NewsAPI[4]. Articles from multiple sources are then displayed on the screen ranging from left-most to right-most view. The selection of news providers is discussed in Section 3. A more technical description is given in Section 6.

[2] `https://fact-checker.herokuapp.com`

[3] `https://www.twitter.com`

[4] `https://newsapi.org`

5 Use Cases and Examples

In this section, we give examples of use cases for our solution. We describe briefly why would one use such a tool and what he or she can gain from it. Then, we show a few examples of what can be expected from it.

5.1 Use cases

First and foremost, we think that anybody can find value in this tool, assuming that they are interested in sorting true from false or misinterpreted information. *BalancedView* is useful to quickly check facts because it provides news articles directly coming from trusted sources. Moreover, it gives a balanced view of the coverage of an event across the political spectrum. By doing so, it insures that the whole information will be unbiased.

For the reason stated above, it makes a valuable tool for journalists more particularly. In their line of work, they often need to fact check information. The website is fast and user-friendly, meaning no technical expertise is required.

5.2 Examples

Example 1 The text used is an extract of a tweet from the account @BorisJohnson. The example in Figure 3 shows a one sided view of a particular situation in British politics. By clicking submit, a selection of articles is displayed from across the political spectrum. In this case, we see that there are no articles returned from the pre-selected right news providers. We come back to this in Section 6 when we mention known issues.

Example 2 The second example in Figure 4 shows a tweet from the account @RealDonaldTrump. We can observe that the output on a tweet is highly evocative of a polarising view. The NewsAPI provides a number of relevant articles from across the political spectrum.

5.3 Extracting Meaning from Tweets

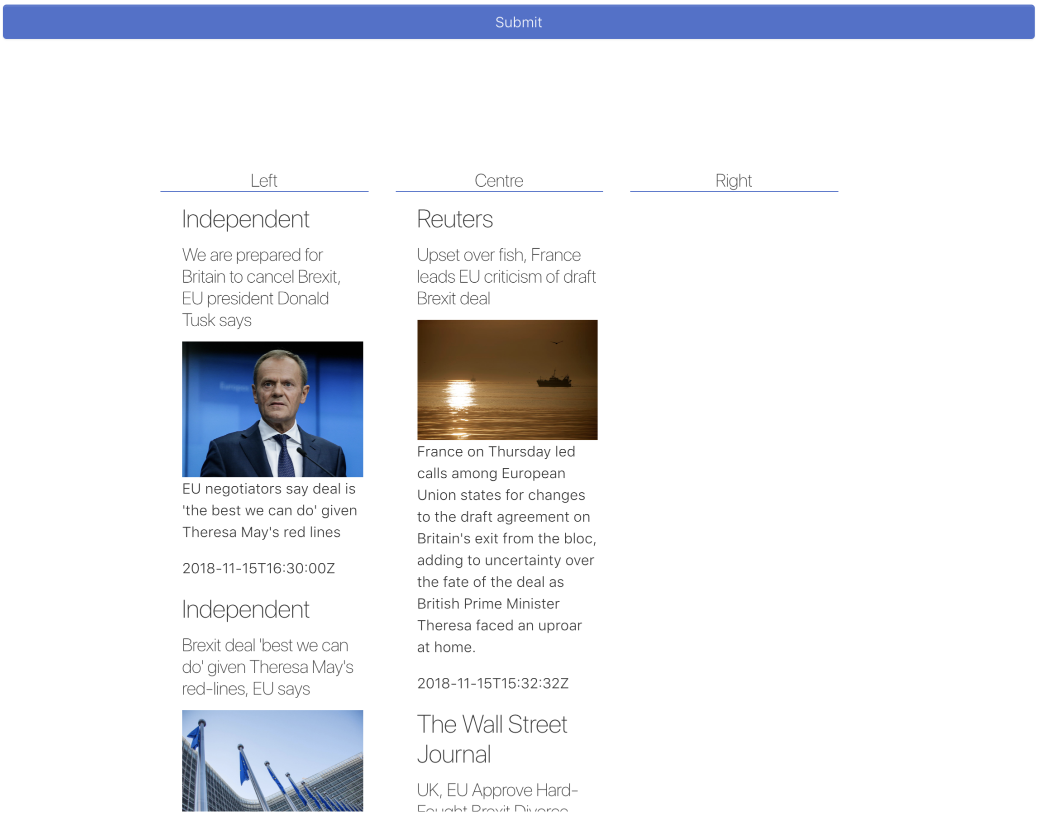
To achieve the initial milestones in this project, the TextRank algorithm is used [7]. TextRank leverages graphical models to extract keywords from text. In particular, an open source Python implementation, *summa* [2] is used, which provides improvements to the seminal algorithm. The keywords extracted from an input text are then used to query the News-API[5]. Requests are sent to a limited number of sources, and displayed according to their political affiliation.

Consider the input text for Figure 3, “Nothing in this political declaration changes the fact that this deal gives the EU a continuing veto over the unilateral power over the entire U.K. to do trade deals or take back control of our laws. We must junk the backstop or we make a nonsense of Brexit.”

A human reader could deduce that the main subject of this text is something to do negotiations in the Brexit process. The TextRank algorithm with key words of three identifies political, declaration, and changes as the keywords. The word Brexit is therefore not used in the query to the News-API. Though, as the result shows, related news is returned.

If the input text is altered to have “Nothing in this Brexit political declaration...” as the first sentence, the results returned differ dramatically, with the keywords becoming Brexit, political, and declaration.

[5] `https://www.newsapi.org`



Independent

We are prepared for Britain to cancel Brexit, EU president Donald Tusk says

EU negotiators say deal is 'the best we can do' given Theresa May's red lines

2018-11-15T16:30:00Z

Independent

Brexit deal 'best we can do' given Theresa May's red-lines, EU says

Reuters

Upset over fish, France leads EU criticism of draft Brexit deal

France on Thursday led calls among European Union states for changes to the draft agreement on Britain's exit from the bloc, adding to uncertainty over the fate of the deal as British Prime Minister Theresa faced an uproar at home.

2018-11-15T15:32:32Z

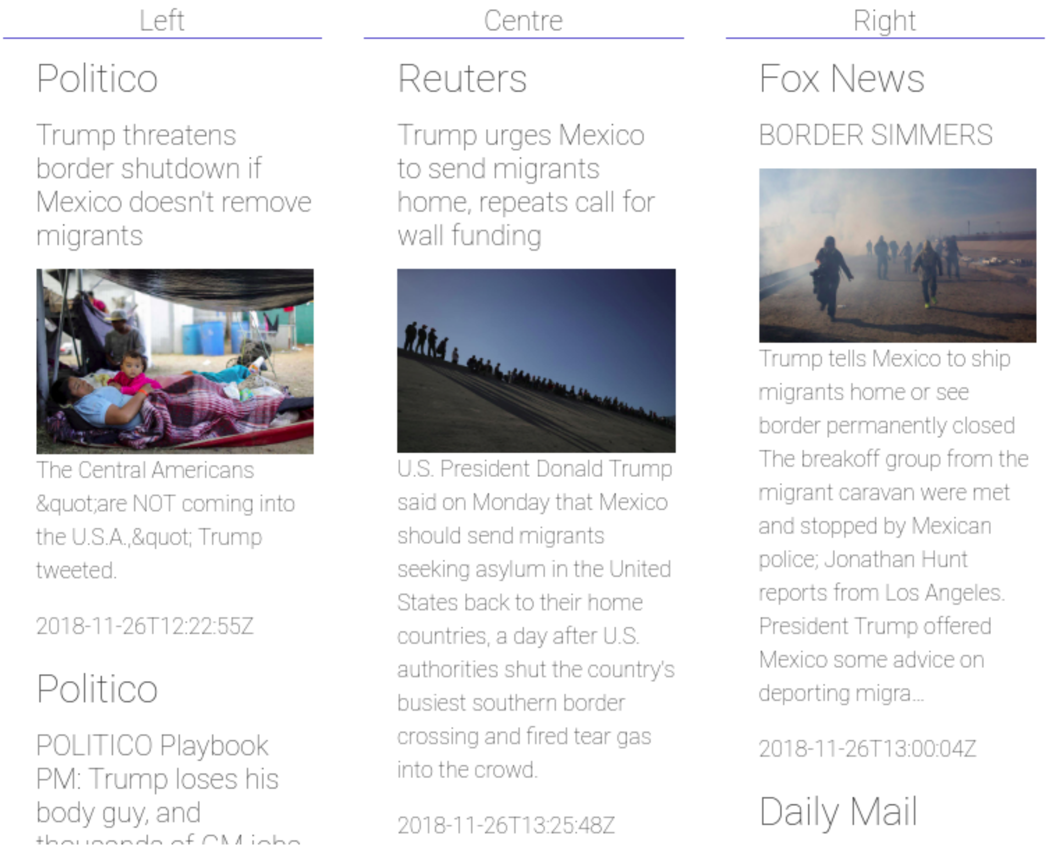
The Wall Street Journal

UK, EU Approve Hard-Fought Brexit Divorce

Figure 3: **Original text**: "Nothing in this political declaration changes the fact that this deal gives the EU a continuing veto over the unilateral power over the entire U.K. to do trade deals or take back control of our laws. We must junk the backstop or we make a nonsense of Brexit." **Author**: @BorisJohnson **Search scope**: 19/11/2018 to 26/11/2018 **Keywords**: political declaration changes

The identified reason for this is the way which the current text extraction places emphasis on certain words and the location of words in the tweet. The current approach limits the number of key words to three and some emphasis is given to words that appear earlier in the tweet. One solution to this would be to increase the number of key words used in building the query. However, some testing was made of this for varying number of words. Increasing the number of keywords beyond four was shown to decrease the likelihood of relevant articles being returned. The reason for this has not been fully explored.

Another solution would be to run multiple queries in the event that no results are returned from the API. For example, generating a second query with an increased number of keywords or a step-wise approach of combining the key words.



Left

Politico

Trump threatens border shutdown if Mexico doesn't remove migrants

The Central Americans &quot;are NOT coming into the U.S.A.,&quot; Trump tweeted.

2018-11-26T12:22:55Z

Politico

POLITICO Playbook PM: Trump loses his body guy, and thousands of GM jobs

Centre

Reuters

Trump urges Mexico to send migrants home, repeats call for wall funding

U.S. President Donald Trump said on Monday that Mexico should send migrants seeking asylum in the United States back to their home countries, a day after U.S. authorities shut the country's busiest southern border crossing and fired tear gas into the crowd.

2018-11-26T13:25:48Z

Right

Fox News

BORDER SIMMERS

Trump tells Mexico to ship migrants home or see border permanently closed The breakoff group from the migrant caravan were met and stopped by Mexican police; Jonathan Hunt reports from Los Angeles. President Trump offered Mexico some advice on deporting migra...

2018-11-26T13:00:04Z

Daily Mail

Figure 4: **Original text**: "Mexico should move the flag waving Migrants, many of whom are stone cold criminals, back to their countries. Do it by plane, do it by bus, do it anyway you want, but they are NOT coming into the U.S.A. We will close the Border permanently if need be. Congress, fund the WALL!" **Author**: @RealDonaldTrump **Search scope**: 19/11/2018 to 26/11/2018 **Keywords**: flag waving migrants

6 Technical Design

In this section, we discuss how the system works under the hood. We will first explain the workings of the technical stack. Then, we detail known problems and potential solutions.

6.1 Overview

As discussed in the previous sections, we propose a system that takes a text from an input and displays a series of articles, sorted by relevance and by political affiliation. We have separated this process into three main steps. Summarising the input, a query to the news providers, display of the results by categories.

6.1.1 Summarising the input

In order to be able to query news providers, it is necessary to summarise the input and extract only the keywords. Among the relevant algorithms, TextRank algorithm and variants of TextRank provide a strong method for doing so [7, 2]. TextRank provides unsupervised identification of centrality of text, using pre-trained algorithms for the low-level tasks like part-of-speech tagging and stemming, as well as graph-based models for the identification of relevant entities.

When the algorithm receives an input, it splits the text and cleans it by removing stop words, numbers and punctuation as well as Twitter's specific keywords such as hash tags and user mentions. The remaining words are going through a part-of-speech filter and only the nouns, adjectives and verbs are kept. Then Porter's stemmer [11] is used to generalise the words further.

From there, the algorithms builds a graph based where each token is a node and edges represent the relations between them. An edge between two words denotes that these two words follow each other in the text. A scoring function assign scores to each node based on the nodes that are reachable from the first word of the input text. In other words, any words for which a path can be found from the starting node will score high. Consequently, words that occur repeatedly or that occur after such repeated words are more likely to score high and words that occur only once at the end of the input will score low.

The keywords are then sorted by decreasing score and the three to five best keywords are kept for the next step. The selection is based on a minimum score of 10%. Both the optimal number of keywords and the minimum score were selected based on the quality and quantity of results after querying the source providers. The whole process described above is depicted in Figure 6.

6.1.2 Querying the news providers

Having identified the keywords, a query is built and sent to the NewsAPI [6], a service that allows us to query a plethora of sources at the same time and get results from a number of countries in multiple languages. However, the free version does not allow to go back more than one month in the past, which limits the presentable results. The sources selection is explained in Section 3.1 and the sources are listed in Table 1.

6.1.3 Displaying the results

As we have discussed in the previous sections, the nudge must be subtle and cannot overload the user with too much information. Consequently, the design must be clean: only the two most relevant articles

[6] `https://newsapi.org`

for each political affiliation are included and only an abstract of the articles is displayed, together with a photo when one is available.

6.2 Known Issues

Importance of the first words TextRank has not been designed to work on small texts like tweets, which are limited to a certain number of characters. We have seen that the algorithm gives more importance to words for which a path can be build starting at the first word of the graph. Consequently, the words that come first take much more importance in comparison to the words that come later in the text, even more so considering that words are not often repeated in tweets. That being said, Twitter users are accustomed to summarise their thoughts in few words and the important words often come up in the front of the text. This can be seen in Figure 6.

Spelling errors and abbreviations This is again a typical problem of tweets. Users often do not write full words and sentences are seldom syntactically correct. TextRank uses part-of-speech tagging to identify words of relevance, which relies mostly on known text corpora and structured texts. If there are spelling mistakes or abbreviations, the words will not be recognised and the results can be disappointing.

Time-limited results As *BalancedView* is currently using the free version of an API to query the news providers, certain limitation are placed on full usability of the API. Primarily, it is only possible to gather results from the previous month. That means that if a user inputs a text that refer to events that happened in the past, the results, if any, will probably not be relevant.

6.3 Potential Improvements

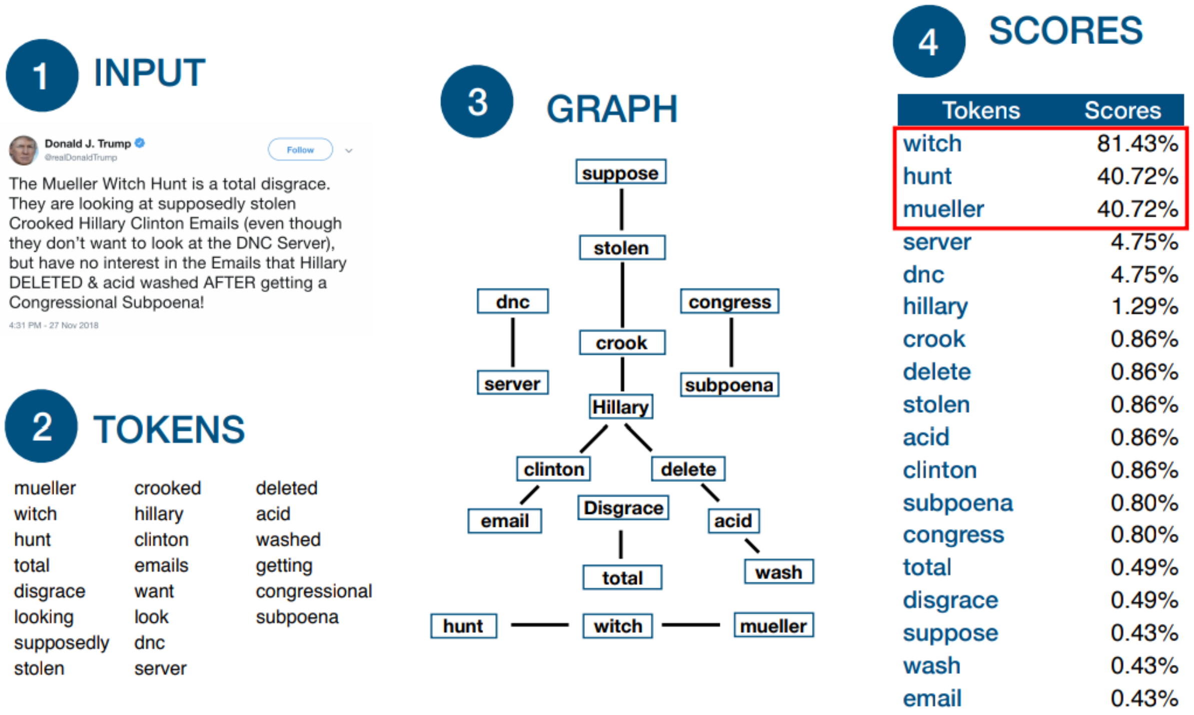
Multilingual Support In the future, it would be possible to adapt the website to accept and work with multi-language input and output, including news providers from multiple countries. However, within the scope of this project, it has not been possible to fully research news providers and their political affiliations in other countries than in the U.S and U.K.

Cross-country and cross-language results It is apparent that the tool could be used to provide results that put into perspective news reported by politically unstable countries by showing articles from other countries, and possibly in other languages, thus offering a solution to state controlled news providers. However, this assumes that the tool is not blocked by governments in said countries.

6.4 Technical Aspects

The application is entirely built using Flask's web framework and Python. In order to provide portability across all operating systems, the app is enclosed in a docker container. The code is publicly available on GitHub and the website is currently hosted on a free Heroku webserver.

Figure 5: *BalancedView*: Technological Stack



Donald J. Trump
@realDonaldTrump
Follow

The Mueller Witch Hunt is a total disgrace. They are looking at supposedly stolen Crooked Hillary Clinton Emails (even though they don't want to look at the DNC Server), but have no interest in the Emails that Hillary DELETED & acid washed AFTER getting a Congressional Subpoena!

4:31 PM - 27 Nov 2018

mueller
witch
hunt
total
disgrace
looking
supposedly
stolen
crooked
hillary
clinton
emails
want
look
dnc
server
deleted
acid
washed
getting
congressional
subpoena

Tokens	Scores
witch	81.43%
hunt	40.72%
mueller	40.72%
server	4.75%
dnc	4.75%
hillary	1.29%
crook	0.86%
delete	0.86%
stolen	0.86%
acid	0.86%
clinton	0.86%
subpoena	0.80%
congress	0.80%
total	0.49%
disgrace	0.49%
suppose	0.43%
wash	0.43%
email	0.43%

Figure 6: **Graph construction**: (1) The input text from Twitter. (2) Token extraction based on the Porter stemmer[11]. (3) Graph is constructed where nodes are words and edges denote whether two words are juxtapositioned in the text (4) Scores reflect whether a node is reachable from the start of the input text. If more than 5 words reach a score of 10%, only the best 5 keywords are selected. If less than 3 keywords reach this threshold, the 3 best keywords are selected.

7 Experiments

7.1 Relevance of the News Articles Presented

A test the ability of the current approach of extracting meaning from tweets. The test measured the the relevance of the articles returned by the tool to the story or situation in the query text.

Method A query was made for each tweet in a set of thirty five tweets covering a number of stories is American and British politics. A trial was a success if at least two of the articles presented first in the result were deemed important to the news surrounding the query.

Results From the 35 trials, the number of successes was counted as eighteen counts of all three articles being relevant, eleven in which only two of three were relevant, and six cases where one or fewer relevant articles were returned. This gives a total of 29 out of 35.

7.2 Effectiveness of the Nudge

An experiment was made to test the usefulness of nudging by presentation and information in the context of perception of news. These experiments were made in survey form, in which participants were presented with Tweets and asked to rate them on either impartiality of trustworthiness.

Method

A survey was made and put to a control group to test whether the nudge was effective in lowering trust in an intentionally selected politically biased tweet. Furthermore, the survey questioned whether the feature of a visual is useful in encouraging users to question reputability of a source.

1. Do people trust obviously unreputable new sources in the absence of a nudge by information? Here a participant is presented with news from a unreputable news source on a news story that does not provoke an emotional response
2. Do people trust obviously reputable news sources in the presence of a nudge by information? A participant is presented with a story from a highly reputable news source, such as the BBC
3. Do people trust news sources of questionable reputability in the presence of a nudge by information? A participant is presented with a story from a verified news source that is unlikely to be known as reputable or unreputable
4. Do people consider politically bias information a fair representation of a view, in the absence of a nudge by presentation?
5. Do people consider politically bias information a fair representation of a view, given a nudge by presentation? The participant is presented with a politically biased statement and linked article, in the presence of the nudge designed in this work

Results

7.2.1 Do people trust obviously untrustworthy news sources in the absence of a nudge by information?

In this case, trust for the news source was low. Responses were concentrated in showing distrust or severe distrust of the news source. However, fifteen percent of respondents placed moderate to high trust in the source despite no verification of the account.

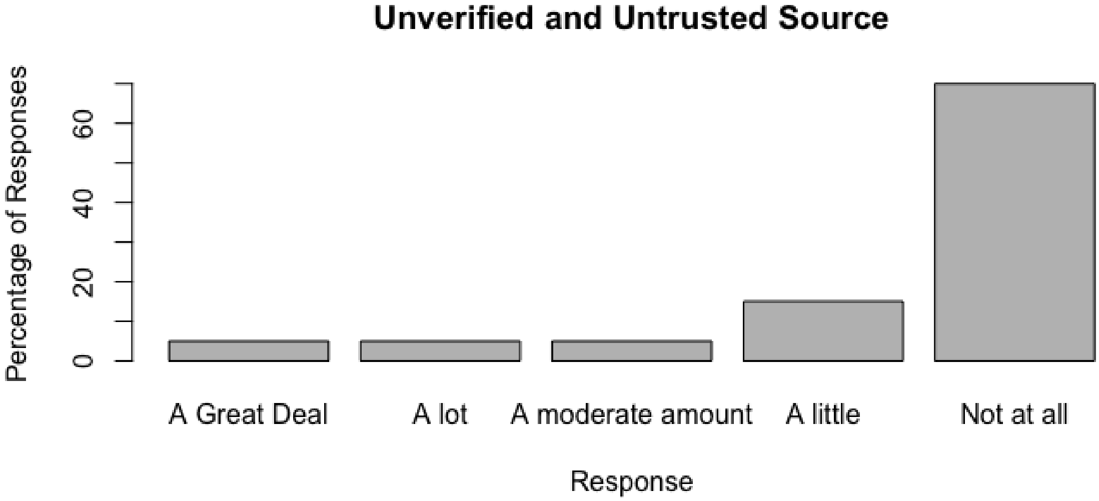


Figure 7: Do people trust obviously untrustworthy news sources in the absence of a nudge by information?

7.2.2 Do people trust obviously reputable news sources?

There was a positive result for this test, people generally tended to trust or highly trust these sources. All respondents trusted the source moderately to highly.

7.2.3 Do people trust news sources of questionable reputability in the presence of a nudge by information?

The results for this test were evenly spread between trusting and not trusting the source. The account was verified and the article featured was produced by a reputable news outlet. The spread of responses shows more trust than in the case of the obviously unreputable source, however less trust is evident than in the case of the highly reputable source.

7.2.4 Do people consider politically bias information a fair representation of a view, given a nudge by presentation? Do people consider politically bias information a fair representation of a view, in the absence of a nudge by presentation?

The key test of the effectiveness of the balanced view nudge is the change in results for questions four and five. From the limited sample size, there is a visible shift in the responses to placing less trust in the singular view.

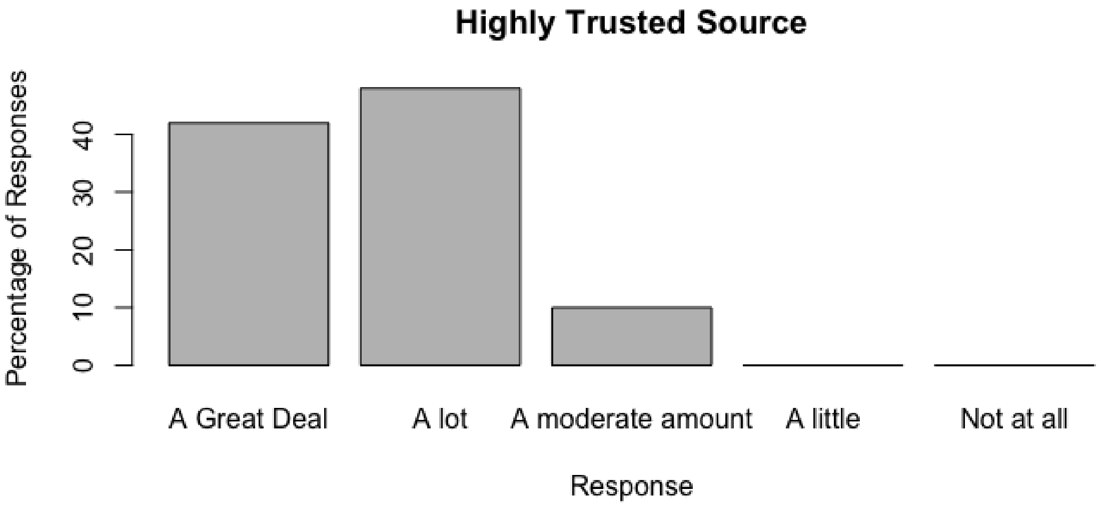


Figure 8: Do people trust obviously reputable news sources?

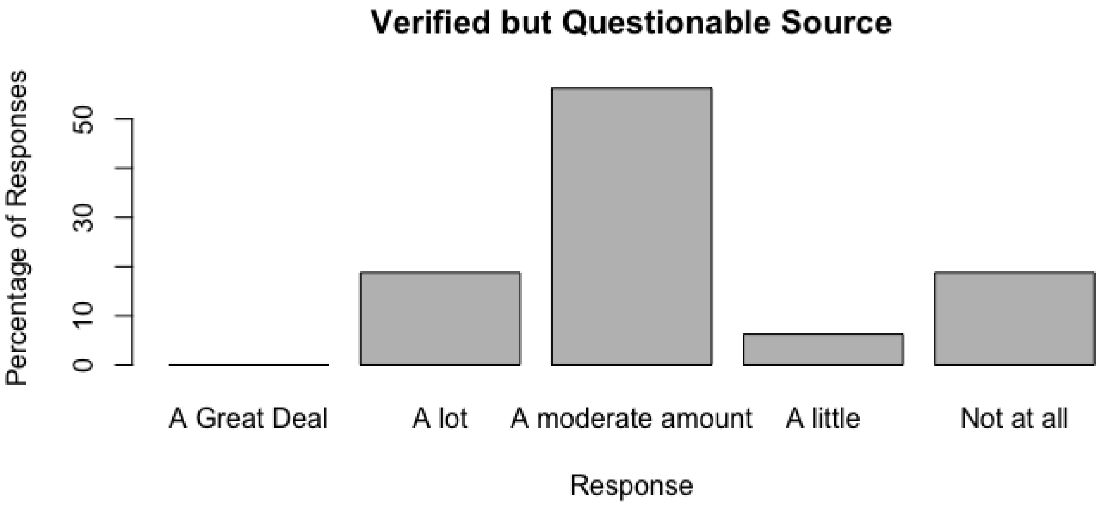


Figure 9: Do people trust news sources of questionable reputability in the presence of a nudge by information?

Initially, positive trust was placed in the fairness of the view being given. In the presence of the nudge, this opinion changed. In this case, results showed that people were generally thought the view was unbalanced.

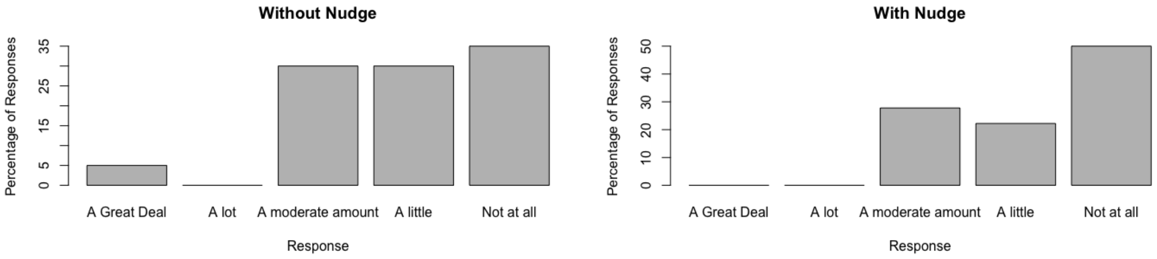


Figure 10: Nudge Comparison

Conclusion and Future Work

We have discussed the problems that are at the root of the spread of misinformation online. Echo chambers and confirmation bias encourage singular views and biased opinions, worsened by the algorithms of platforms and their action to target content. The combination of these factors creates a feedback system that allows the propagation of unreliable news.

We have discussed nudges as a solution to this. This involves gently steering users towards adopting fact checking habits in their behaviour online. Two nudging strategies were proposed: one that presents results in a way that pushes the user to look further and another that gives feedback on the quality of the posts that are shared online. The former option was implemented into an online tool that can be used to quickly browse articles relating to information expressed in a short text such as a tweet. The articles come from trustworthy news providers and are classified into political categories. Simply, the tool can be used to quickly and efficiently fact check any piece of information that one might read online.

Even though the second nudging strategy was not implemented, we have compared its effectiveness with our tool by the means of an online survey. The objective of the experiment was to assess how questionable article were perceived without any nudging strategy and with one of the two approaches discussed. The results show that the nudging strategies makes people more aware of the trustworthiness of the sources. Furthermore, there is potential in presenting a balanced view of related news as a solution to lowering acceptance of a singular view.

Future Work

These findings are encouraging. However, work remains to be done in a number of areas. Existing issues in Section 6 as well as potential improvements must be investigated. Furthermore, the number of sources available should be increased. This expansion of sources could reflect a deeper range of of political views. To enable this, there is a need for researchers studying the trustworthiness of sources along the political spectrum. Although all sources could be included, some judgement has to be maintained be made to ensure journalistic integrity.

An issue that has been discussed is the relative efficiency of *TextRank* algorithm. In this regard, recent advances in Deep Learning have uncovered the relevancy of neural networks for processing text. For example, attention mechanisms allow to focus on some parts of the texts that are more relevant to the task at hand. It could be interesting to investigate further whether such models can be of use to efficiently extract relevant keywords from tweets.

Another direction for development is to investigate a number of other languages both to process the input text and to provide multilingual articles. Currently the news sources are British and American, and only in English. Developing to include French, Dutch and Spanish would give considerable scope for trials in areas in which fake news and polarised politics are evident on online platforms.

Finally, limited experimentation has been undertaken and the full measurement of the effectiveness of the nudge remains to be shown. This can be achieved through recording interactions in the user experience, for example measuring how much time the users spend on the page and if they still share and propagate unreliable news after having been in contact with *BalancedView*.

References

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. Nudges for privacy and security: understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)*, 50(3):44, 2017.

[2] Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606, 2016. URL `http://arxiv.org/abs/1602.03606`.

[3] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. *arXiv preprint arXiv:1801.01665*, 2018.

[4] Silvia Knobloch-Westerwick and Steven B Kleinman. Preelection selective exposure: Confirmation bias versus informational utility. *Communication Research*, 39(2):170–193, 2012.

[5] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[6] D.J. Levitin. *A Field Guide to Lies and Statistics*. Penguin Books Limited (UK), 2018. ISBN 9780241974872. URL `https://books.google.be/books?id=RhtDMQAACAAJ`.

[7] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004.

[8] Tobias Mirsch, Christiane Lehrer, and Reinhard Jung. Digital nudging: altering user behavior in digital environments. *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017)*, pages 634–648, 2017.

[9] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. Political polarization & media habits. *Pew Research Center*, 21, 2014.

[10] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175, 1998.

[11] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980. doi: 10.1108/eb046814. URL `http://www.emeraldinsight.com/doi/abs/10.1108/eb046814`.

[12] Matthew Smith. How left or right-wing are the uk's newspapers?, 2017. Available at: `https://yougov.co.uk/topics/media/articles-reports/2017/03/07/how-left-or-right-wing-are-uks-newspapers` Last accessed: 27/11/2018.

[13] Richard H Thaler. Nudge: Improving decisions about health, wealth, and happiness, 2008.

[14] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. A field trial of privacy nudges for facebook. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2367–2376. ACM, 2014.