# Social Media as an Auxiliary News Source

Stephen Bradshaw[1], Colm O'Riordan[1] and Riad Cheikh[2]
[1] *School of Computer Science, National University Ireland Galway, Ireland*
[2] *South Gloucestershire and Stroud College, U.K.*

Keywords: Data Mining, Information Extraction.

Abstract: Obtaining a balanced view of an issue can be a time consuming and arduous task. A reader using only one source of information is in danger of being exposed to an author's particular slant on a given issue. For many events, social media provides a range of expressions and views on a topic. In this paper we explore the feasibility of mining alternative data and information-sources to better inform users on the issues associated with a topic. For the purpose of gauging the feasibility of augmenting available content with related information, a text similarity metric is adopted to measure relevance of the auxiliary text. The developed system extracts related content from two distinct social media sources, Reddit and Twitter. The results are evaluated through conducting a user survey on the relevance of the returned results. A two tailed Wilcoxon test is applied to evaluate the relevance of addition information snippets. Our results show that by partaking the experiment a users' level of awareness is augmented, second, that it is possible to better inform the user with information extract from a online microblogging sites.

## 1 INTRODUCTION

News articles are a long-established source for informing oneself on a topic. However an article is typically written by a single author or a limited set of authors, so naturally may be biased towards that author's viewpoint. In addition, a newspaper company might have a vested interest in an issue and as such, frame a story to a particular slant to match their own views. In the referendum in Scotland regarding their proposed cessation from the United Kingdom (2014), there were many claims leveraged by the *yes* campaign, which stated that there were no pro-cessation articles to be found in any of the national news outlets (Monbiot, 2014). Indeed, this accusation has once more been voiced in the more recent (2016) debate on whether Britain should leave the EU (Osborne, 2016). It is argued here that aspects of web 2.0 can be mined for knowledge, that will combat bias and polarisation found in main stream media, to gain a more in-depth understanding of an issue.

In this paper we will engage with the following hypothesis, *Social Media is a rich source of information which, when appropriately mined can be used to augment an existing knowledge snippet.* To gain a better intuition of this hypothesis we propose two research questions:

- *RSQ1* Given a source of information, can we show that additional data can be mined from social media, which is relevant to the original information snippet.
- *RSQ2* Can we determine if one source is more appropriate for extracting a relevant information snippet than the other.

There are many sources of information available to users when looking to inform themselves on a topic; some of these include news articles, blogs, Wikipedia (an open source community created encyclopedia) and Twitter (a popular microblogging site). News articles and Wikipedia can be great sources of information, however, they tend to be written by sole authors or, by a small set of authors, and as such may suffer from the author's bias. Additionally, users have a tendency to read content with which they agree. This is a much studied concept particularly in the field of recommender systems and is discussed under a number of headings, including the *filter bubble* (Pariser, 2011), *echo chamber* (Colleoni et al., 2014) and *balkanization* (Cosley et al., 2003). In short, these are different terminology for the same issue, that content based recommendation approaches are prone to recommending content that is overly similar to what has already been read. The challenge then is finding content that is diverse while still being of interest to the user. Social chat forums can be a great source of differing opinion, though it can be a time consuming process to read through all of the threads.

The aim for what follows is to explore the fea-

Bradshaw, S., O'Riordan, C. and Cheikh, R.
Social Media as an Auxiliary News Source.
DOI: 10.5220/0010144202770281
In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) - Volume 1: KDIR*, pages 277-281
ISBN: 978-989-758-474-9

sibility of improving the informative value of text, through sourcing additional information from social forums. Additionally, we look to determine if Reddit or Twitter is a superior source of information for augmenting information snippets. In this paper, using extracts from an article regarding a recent news event, we compare Reddit and Twitter as alternative sources of auxiliary information and measure the relevance of comments found there to a set of users. Text similarity is applied to a body of text and a selection of microblogs to link comments from Reddit and tweets from Twitter with paragraphs written on a particular topic. Subsequently, a group survey is performed to evaluate the results; a statistical analysis is then applied on the user feedback to determine which were significant. Our conclusions show that relevant information can be identified which improves the dataset; additionally, the results show that Twitter proved to be a more appropriate source. Included will be a number of explanations as to why this might be the case.

## 2 RELATED WORK

Hsu et al. (Hsu et al., 2009) present a study which aims to rank comments found on articles in *Digg* (a popular online opinion editorial). They apply logistic regression using a selection of metadata found relating to the comments and the social connectedness of the comment makers. Interestingly, they found that visibility is one of the most influential factors in how a comment is rated. Visibility is affected by how quickly one responds to an article. Visibility can be influenced by a number of factors external to the actual content, such as time of posting or thread bias (Hsu et al., 2009). By focusing on the text alone, one can then promote content that is relevant, rather than content that is dependent on up-votes or retweets to gain visibility.

Shmueli et al. (Shmueli et al., 2012) use co-commenting and textual tags to implement a collaborative filtering model which recommends news articles to users. Past comments and social connectedness are used as indicators for recommendations. Others who have investigated this approach include work by Li (Li et al., 2010) and Bach (Bach et al., 2016). Such an approach lends itself well for identifying related content of interest, though it does not address other issues such as problems associated with the echo-chamber.

Another approach proposed in the literature clusters trending tweets together and extracts common terms (Rehem et al., 2016). These terms are used to form a query which the authors use to enter into a search engine. They extract the most related news articles to the query and recommend them to the user. As a recommender system this approach is not very personalised (as no input is made of user tastes), it is based on the assumption that popular news is of interest to the user. It does however add a level of serendipity and helps to combat negative effects such as the filter bubble by not over-fitting to the user's tastes.

A similar approach to the one outlined in this paper, is presented by Aker et al,. (Aker et al., 2016) who aim to link comments to articles. They linked comments to articles at a sentence level, where the article is first reduced to sentences and each sentence is then compared to the related comments. Their dataset was constructed using articles published online in the Guardian newspaper website between the months of June-July 2014. They amassed a total of 3,362 articles with an average of 425.95 comments per article. In addition to linking comments and sentences, they perform sentiment analysis, to determine whether the comments agreed or disagreed with a given article. As distance metrics they employed Jaccard and Cosine distance. As well as using syntactical similarities they analysed connections using distributional similarities between terms. Distributional similarity is the idea that words that co-occur regularly have a similarity of meaning. Using two additional sources (BNC corpora and Wikipedia) they constructed similarity vectors and incorporated those into their system. They include on average about 425 comments per article. One limitation of their approach is that they rely on manual labelling to determine polarity in the debate, which is an acceptable early first step. Ultimately, a functional system would have to identify indicators automatically for any degree of practical application. Bias polarity could potentially be inferred from the sources used in this experiment, by determining poster demographic in a given subreddit, or through looking at additional tweets from the individuals who retweet a given text.

Research by Becker (Becker et al., 2010) utilises comments found in *flickr* to identify events, as determined from observing active conversation found on social media. The authors identify distinctions between event detection in social media with event detection in more standard datasets; namely, that there is less structure and more noise in social media data. They cluster comments into related groups and from the resulting clusters deduce if a particular event is happening. Similarly, a lot of work in this field focuses on summarising user comments rather than mapping them to points made in any corresponding news article (Ma et al., 2012) (Hu et al., 2008) (Khabiri et al., 2011).

Table 1: Top User Relevance Feedback.

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | Avg Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Reddit | 3 | 4 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 2 | 2.9 |
| Twitter | 4 | 2 | 4 | 4 | 2 | 4 | 2 | 3 | 3 | 3 | 3.1 |

# 3 DATASET

The data used in the experiments is taken from three different sources. The news article is taken from the BBC homepage which is a comprehensive summary the issues associated with the impact of Brexit [1]. Additionally, 25,000 tweets from 19/07/2016 to 29/07/2016 were collected, comprising tweets that featured the term or hashtag *Brexit*. The timescale is such, that it encompasses tweets from before and after the referendum. Finally, a collection of Reddit comments was gathered from a subreddit called */r/brexit* and is composed of 20,000 discussion posts on the issue.

Each paragraph in the news article is processed and represented as a term vector. Similarly, each tweet and Reddit comment is also represented in the same fashion. A TF-IDF weighting scheme is applied to determine a value for the terms. A cosine similarity metric is applied to each comment (taken from Twitter and Reddit) to determine how similar it is to each particular paragraph from the news article. Selecting the first ten paragraphs that contain 80 or more words, they are presented to the user with the top five ranked comments from each alternative source. The user is then given the opportunity to rate how relevant the comments/tweets are to each paragraph. User rankings range from one to five, indicating totally non-relevant to extremely relevant respectively. Subsequently, the values of the top five comments for each paragraph, from both the Twitter and Reddit sources are measured. A table is plotted from the resulting scores; a statistical analysis is then performed on the top comments of each set to gain a better perspective on the performance of the approach.

# 4 METHODOLOGY

Initial steps involved obtaining an article from the BBC which was felt to give a robust insight on the issue. It contained seventy-six paragraphs and had a header for each topic, followed by an explanation on what the topic was and how it impacted the debate. Ten of these descriptive paragraphs were selected for the experiments. For each paragraph, the similarity level was compared with comments from the respective datasets (Twitter and Reddit). The top five most related comments/tweets in order of similarity were returned to be evaluated by the user.

Evaluation was determined through user feedback of relevance; 23 students were invited to interact with our system. Each user was presented with a news paragraph and 5 related comments from each domain. Users were then asked to rate the relatedness of each comment presented on a scale of one for not related and five for extremely related. Their feedback is analysed using the Wilcoxon statistical test.

# 5 RESULTS

## 5.1 Analysing Top Comments

In Table 1, we present the data indicating the user relevance rankings of the top returned tweets and Reddit comments. This table represents the top score of most related comment from each domain. In the majority of cases, the relevance found is quite good, showing that we can augment existing material with related comments from social media forums.

The overall relevance of all five comments to the paragraphs is displayed in Table 2. The data shows that Twitter instances are more relevant in all but one of the cases. Based on user feedback we conclude that there are a number of reasons for this. First, the nature of Twitter comments is that they are more self-contained. This is to say that they are written with the intent of standing alone. This is in contrast to the Reddit comments, each of which are an addition to an ongoing conversation. Second, the Reddit comments contain more anaphors; which adds an element of ambiguity to the target of the conversation. Third, many Twitter comments contain links to additional articles. At this stage of the project we did not parse these external sources, the participants however reported that the presence of a link reinforced the validity for opinions expressed.

[1] https://www.bbc.com/news/uk-politics-32810887

Table 2: Sum Average Comparison of Average User Relevance Feedback.

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | Sum Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Reddit | 17 | 18 | 13 | 17 | 11 | 15 | 11 | 14 | 15 | 9 | 140 |
| Twitter | 16 | 15 | 18 | 18 | 15 | 17 | 10 | 16 | 17 | 16 | 158 |

Table 3: User Informativeness Feedback.

| More Informed | Equally Informed | Less Informed |
|---|---|---|
| 9 | 10 | 4 |

## 5.2 Statistical Analysis

To gain a greater insight into the data we investigated how the top comment from each domain compare. The intuition behind this is that people are often more interested in the top result, and interest levels diminish the further down a list one goes. Table 1 shows a summary of the user feedback in relation to the top comment from Reddit and Twitter. Twitter has more top ranked comments than Reddit, having a higher relevance score in 4 of the 10 paragraphs.

Table 2 presents a sum total of all comments. We performed a Wilcoxon t-test on the figures testing the hypothesis that Reddit is a superior source of information for obtaining additional related data (see table 4). This analysis included feedback on all comment relevance scores per paragraph. Only comment 1 is a significant result supporting this claim. Conversely, of the seven instances when the feedback from the Twitter exceeded that of Reddit, three are significantly different. Thus we can conclude that Twitter performed significantly better as an additional source of information. It is clear from looking at this table that the intuition provided from the Wilcoxon test are accurate. The feedback given in relation to P2 are discernibly in favour of Reddit, while P3, P4, and P10 are substantially weighted in favour of Twitter.

Finally, we asked participants to assess their knowledge of the situation prior to conducting the survey and then again after. Table 3 contains a breakdown of responses. In nine of the instances, users reported that they felt more informed on the topic then when they had started, 10 users reported an equal level of awareness, while 4 reported that their knowledge of the topic was lower having completed the survey. This result was unusual but on follow up it was determined that their awareness of the topic expanded as result of the survey, such that they re-assessed how much they were actually aware of the information. This phenomenon is known as the *Dunning Kruger effect* and is a well documented concept in psychology (Dunning et al., 2003).

# 6 CONCLUSIONS

We set out to determine if useful additional comments/tweets can be source to improve upon existing factual representation of an argument. We conducted a user study and evaluated the results. Our findings are that 1) social media has the potential to be a relevant source of additional information, and 2) in these initial experiments, Twitter showed itself to be a more informative source of information than Reddit. Though the average relevance score found from both sources showed that either could be considered as viable auxiliary sources. It should be noted that the relative small scale of the survey population (23), is not sufficiently large to determine that one social media based source is superior to another, just that given that population size it has shown itself to be more relevant. The survey more, acts as an opening salvo in the investigation of linking social media based content with that of official sources. The experiment asks more questions than it answers; from this we propose that there are three further avenues of research.

First, the selection of related comments was based on a similarity metric. This can be improved through using a clustering approach as a preprocessing step. Namely, clustering the related comments to identify prevalent topics expressed, and applying a frequency metric such that, the returned recommended comments are a) relevant and b) reflective of commonly asserted opinions. Additionally one could consider alternative clustering approaches such as K-nearest neighbour and an exploration on the impact of using different text similarity approaches; such as Jaccard Similarity Coefficient or Manhattan Distance.

Second, follow-up work will include alternative methods for linking content with data and more extensive evaluations over more material. Greater linkage could potentially be achieved through employing a word sense disambiguation step that would use word embeddings to resolve intend use of terms (Bradshaw et al., 2017). This could then be used to more accurately inform relatedness found in the suggested information snippets.

Finally, information related to the paragraphs themselves can be further extrapolated through, analysing the related number of comments to a given paragraph as well as the general sentiment expressed in these comment.

Table 4: Shows results from Wilcoxon Test.

| Difference | N | Statistic | P-Value | Reddit mean | Twitter mean |
|---|---|---|---|---|---|
| T1 - R1 | 23 | -1.55 | .12 | 3.04 | 3.65 |
| T2 - R2 | 23 | -3.39 | .001 | 3.78 | 2.35 |
| T3 - R3 | 23 | -3.57 | .000 | 3.30 | 4.30 |
| T4 - R4 | 23 | -2.85 | .004 | 3.22 | 4.17 |
| T5 - R5 | 23 | -1.31 | .190 | 1.70 | 1.91 |
| T6 - R6 | 23 | -1.59 | .111 | 3.09 | 3.70 |
| T7 - R7 | 23 | -2.29 | .022 | 2.87 | 2.00 |
| T8 - R8 | 23 | -.19 | .851 | 3.00 | 3.00 |
| T9 - R9 | 23 | -1.48 | .138 | 3.09 | 2.78 |
| T10 - R10 | 23 | -3.49 | .000 | 1.83 | 3.30 |

## REFERENCES

Aker, A., Celli, F., Kurtic, E., and Gaizauskas, R. (2016). Sheffield-trento system for comment-to-article linking and argument structure annotation in the online news domain.

Bach, N. X., Do Hai, N., and Phuong, T. M. (2016). Personalized recommendation of stories for commenting in forum-based social media. *Information Sciences*, 352:48–60.

Becker, H., Naaman, M., and Gravano, L. (2010). Learning similarity metrics for event identification in social media. In *Proceedings of the third ACM international conference on Web search and data mining*.

Bradshaw, S., O'Riordan, C., and Bradshaw, D. (2017). Constructing language models from online forms to aid better document representation for more effective clustering. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, pages 67–81. Springer.

Colleoni, E., Rozza, A., and Arvidsson, A. (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332.

Cosley, D., Ludford, P., and Terveen, L. (2003). Studying the effect of similarity in online task-focused interactions. In *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work*, pages 321–329. ACM.

Dunning, D., Johnson, K., Ehrlinger, J., and Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current directions in psychological science*, 12(3):83–87.

Hsu, C.-F., Khabiri, E., and Caverlee, J. (2009). Ranking comments on the social web. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 90–97. IEEE.

Hu, M., Sun, A., and Lim, E.-P. (2008). Comments-oriented document summarization: understanding documents with readers' feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 291–298. ACM.

Khabiri, E., Caverlee, J., and Hsu, C.-F. (2011). Summarizing user-contributed comments. In *ICWSM*.

Li, Q., Wang, J., Chen, Y. P., and Lin, Z. (2010). User comments for news recommendation in forum-based social media. *Information Sciences*, 180(24):4929–4939.

Ma, Z., Sun, A., Yuan, Q., and Cong, G. (2012). Topic-driven reader comments summarization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 265–274.

Monbiot, G. (2014). How the media shafted the people of scotland. *The Gaurdian*.

Osborne, S. (2016). Eu referendum: National press biased in favour of brexit, says study. *The Gaurdian*.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Rehem, D., Oliveira, J., França, T., Brito, W., and Motta, C. (2016). News recommendation based on tweets for understanding of opinion variation and events. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 1182–1185. ACM.

Shmueli, E., Kagian, A., Koren, Y., and Lempel, R. (2012). Care to comment?: recommendations for commenting on news stories. In *Proceedings of the 21st international conference on World Wide Web*, pages 429–438. ACM.