Towards Automatic Assessment of the Social Media Impact of News Content

Tom De Nies* Gerald Haesendonck* Fréderic Godin*
Wesley De Neve*† Erik Mannens* Rik Van de Walle*
{tom.denies,gerald.haesendonck,frederic.godin,wesley.deneve,erik.mannens,rik.vandewalle}@ugent.be

*Ghent University - iMinds
Department of Electronics and Information Systems, Multimedia Lab
Gaston Crommenlaan 8 bus 201
B-9050 Ledeberg-Ghent, Belgium

†KAIST
Dept. of Electrical Engineering
Image and Video Systems Lab
Republic of Korea

ABSTRACT

In this paper, we investigate the possibilities to estimate the impact the content of a news article has on social media, and in particular on Twitter. We propose an approach that makes use of captured and temporarily stored microposts found in social media, and compares their relevance to an arbitrary news article. These results are used to derive key indicators of the social media impact of the specified content. We describe each step of our approach, provide a first implementation, and discuss the most imminent challenges and discussion points.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval —*Information filtering, Relevance feedback*

Keywords

News; Social Media; Relevance; Content Analysis; Search and Retrieval

1. INTRODUCTION

In today's media landscape, journalists cannot ignore the impact of social media on the content they create. Social media play a large role in how a news item is disseminated, interpreted and searched [12]. Often, a news item will spread on social media even before it is published through a more traditional channel [7, 8]. In fact, it has been shown that although only an estimated 5% of Twitter content is news [15], over 85% of tweets on trending topics can be classified as headline news or persistent news [10]. Therefore, assessment of the social media impact of a news item is a valuable tool in the context of news selection, creation and strategical advertising. However, due to the vast daily throughput of data on social media, and the limitations of a number of social media APIs, such as a limit on the number of API calls per application, this is not a trivial task.

In this paper, we propose a novel method for assessing the social media impact of an arbitrary news text. To do this, we accumulate a stream of microposts for temporary storage and make it available for querying. This way, we aim to find as many microposts related to the news content as possible. Using the statistical properties of these microposts, we present information regarding the social media impact – and thus, the *interestingness* – of the content. This application is primarily targeted towards actors in the publishing industry, such as journalists, editors and advertisers.

The rest of the paper is structured as follows: first, we provide a brief survey of related publications and approaches found in literature in Section 2. Next, we break the proposed approach down in three steps, and describe the requirements and rationale for each step in Section 3. Since the approach is deliberately kept general, we also describe a first, naive implementation to provide additional clarification in Section 4. In Section 5, we provide a number of discussion points. Here we also list the most important challenges encountered and describe the future steps for further development and evaluation of the approach. We end this paper in Section 6 with a conclusion.

2. RELATED WORK

The interplay between social media and news is extensively studied in literature, using very diverse methods. In [13], social dynamics are used to predict the popularity of news content on the social news site Digg.com. Whereas the approach in [13] focuses on predicting the interestingness of an article using statistical knowledge, such as the number of votes the article gets, our approach is complementary, as it considers the content of the article itself. The authors of [14] analyze the content features of quickly spreading items on Twitter. This involves identifying features for interestingness, and insights into what makes a message on Twitter worth retweeting, which is highly related to our approach. In [1], a different use case than news is chosen, namely the predicting of box office revenues for several movies based on the related social media activity, but the principles are comparable to our application.

In [5], several contributions related to the retrieval component of our approach are made, i.e., the generation of complementary summaries of information collected across news and social media streams, and a measure to compare sentences across these media streams. Further investigation into these concepts could certainly prove useful when

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

expanding our approach beyond the first naive implementation, as described in Section 4. The reverse problem is discussed in [16], where the authors propose a method to select relevant content from a news stream, based on the social media information of a user.

When it comes to efficient querying, several approaches have been proposed in the research community [3, 9] and even at Twitter itself [2] to optimize the storage and indexing of a large collection of microposts. Implementing one of these approaches – or an alternative solution – will be imperative for usage of our application in a real-world scenario.

3. PROPOSED APPROACH

Our approach is segmented into three general parts: *storage and querying* of the micropost data, *retrieval* of the relevant microposts, and *presentation* of the results. In Figure 1, we show an architectural overview of our proposed system. In the next sections, we discuss the requirements and desired functions of each component, without narrowing the approach to specific social media or technologies. A more tangible description is provided in Section 4, where an example implementation is described.

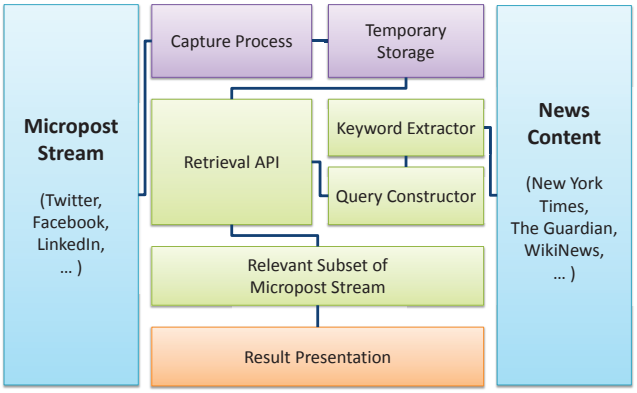

Figure 1: An overview of the components of the proposed approach.

3.1 Micropost Storage and Querying

The first component of our approach ensures the access to the data available on the social media. For our work, we are mainly interested in the social media that incorporate the concept of *microposts*, such as small statements, status updates or quotes. Generally, these social media provide two types of access to their micropost data: either via a *search API*, or a *streaming API* (or both). The main difference is that a search API can provide access to data backward in time, whereas a streaming API functions as a capturing tool, only providing access to data from the moment the user starts capturing. For our use case, a search API would be most suitable, since we do not know in advance what to look for. However, as this is very valuable data for various stakeholders (e.g., business analytics, advertisers, etc.), most social media platforms do not provide unrestricted and/or free access to their search API. Therefore, depending on whether unrestricted access to a search API is available, alternative solutions must be developed to allow temporary storage and querying of the content gathered from a streaming API. This content could be stored only for a limited timespan, as relevance of most microposts decreases over time. However, which posts to store and for how long, is an implementation choice, dependent on the use case.

In any case, an access point to the micropost data must be provided – be it through a search API or an accumulated micropost stream – to facilitate the efficient retrieval of microposts, as described in Section 3.2.

3.2 Retrieval of Relevant Microposts

The second component handles micropost retrieval. The goal of this component is to retrieve posts relevant to an input news article. To achieve this, several techniques can be applied. In Section 2, we list a number of these techniques. While the possibilities are immensely diverse, all solutions have one key functionality: the comparison of microposts to other types of content (in our case: news). Independent of the implementation, the retrieval component must provide the following functionality:

- Allow the content of a news article to be entered as input data;
- process this content, and compare it to the collection of microposts;
- return a list of microposts, sorted by descending relevance to the news article.

These returned results are then analyzed, interpreted, and presented to the user, as described in Section 3.3. Note that the methods and sorting criteria used for querying and ranking of the microposts are modular and interchangeable, and more optimal solutions might be available.

3.3 Interpretation & Presentation

Now that the set of microposts related to the news content is available, we must present information regarding the impact of these posts to the end user. Deciding which information to show the user, and how to derive it, is an empirical process, which will have to be evaluated during use of the approach in practice. However, we are able to discern two minimum levels of abstraction. First, it is important to derive **high-level indicators** of the impact of the news content on the social media. Second, we also need to provide **references** to the original microposts, to ensure reproducibility of the derived indicators, as well as to allow the end-users to draw their own conclusions, based on the data.

Two key indicators that are an essential part of the majority of social media, are *post endorsements* and *user influence*. These metrics indicate whether or not a post was endorsed by a lot of people (e.g., how many time a post was retweeted, shared or liked), and/or whether the person who made the post was very influential (e.g., whether he/she has a lot of followers, friends or connections). An example of a micropost having an exceptional amount of endorsements and a high influence, is the so called "Golden Tweet" [4]. It was authored by Barack Obama, who has over 27.7 million followers on Twitter at the time of writing, and was retweeted 813,481 times in 2012.

In this context, it could also be useful to make use of existing services that measure influence, such as Klout[1]. Klout is specialized in measuring the influence of users of several

[1]http://klout.com

social networks and could be an interesting addition to the system.

4. NAIVE IMPLEMENTATION

Due to its novel nature, the approach proposed in Section 3 is deliberately kept general, to allow for several possible implementation methods. Each of the components has one or more features that can differ for each implementation, allowing a large number of optimizations and design decisions to be made. However, to provide a tangible example of the merits of our approach, we created a first, naive implementation.

For the first implementation, we chose the Twitter Streaming API[2] as the social media API. Aside from following specific persons, keywords or hashtags, this API allows capturing 1% of *all* global microposts on Twitter, which makes it particularly interesting for the first evaluation of our approach. In our case, we are mainly interested in three data fields of a micropost: the text or message, the number of retweets (*post endorsements*) and the number of followers (*user influence*). Note that these fields could also easily be extracted from an additional social media site, such as Facebook. In that case, the post endorsements would be quantified by the number of 'likes' a post receives, whereas the number of friends of a user would signify the user influence. Additionally to these three fields, we also keep the id, the time of creation, and the author of the micropost.

Once the capturing process is started, we transfer all captured data to a flexible storage system, such as a database. For our implementation, we chose to store the microposts in an instance of Apache Solr[3], a database system that supports efficient indexing and querying. To ensure efficient access, indexes were created on all data fields. Additionally, we chose to implement a RESTful interface for easy access to the data. This interface allows the database to be queried using keywords and regular expressions, with several sorting options. We chose to keep the microposts stored for one week, after which they are automatically removed from the database.

To retrieve relevant microposts, we extract a number of keywords from the article, characterizing the semantics of its content. To achieve this, we employed a Named Entity Recognition (NER) service, more specifically AlchemyAPI[4]. This service facilitates the extraction of keywords from the input news content. This service extracts Named Entities (NEs) from plain text, linking each entity to a URI on the web to facilitate disambiguation. Moreover, AlchemyAPI features a service specifically for the extraction of textual keywords. This Keyword Extraction API is applied to the input article. The extracted keywords are then used to form one or more queries, used to retrieve a set of relevant microposts. In our case, the database is queried using each of the extracted keywords, and the results are merged into one list. This set of microposts is then ranked using a sorting criterion giving preference to posts containing more of the extracted keywords. This way, we aim to rank the most relevant microposts among the first results in the list. If there are ties, the tied results are sorted by descending user influence (i.e., followers), and post endorsements (i.e., retweets), respectively.

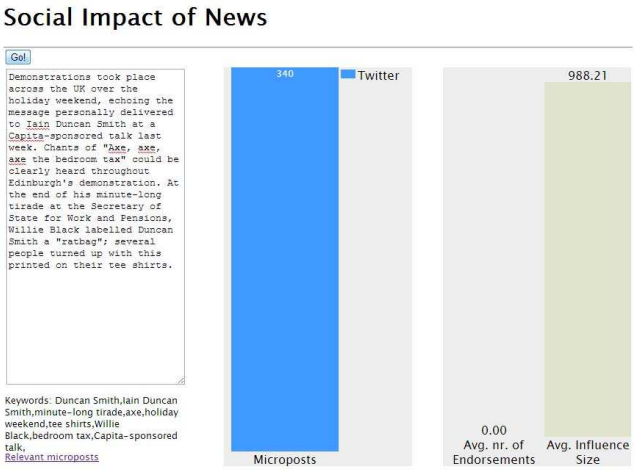

Figure 2: Example of the graphical output of the first implementation of the proposed approach.

The next step is to calculate a global indicator of the levels of influence and endorsement of the most relevant microposts to the input news article. To do this, we consider the N highest ranked – and thus, most relevant – results of the retrieved microposts, with N a user-defined number. Of these N microposts, we calculate the average number of endorsements and user influence. Finally, we show the following information to the user:

- total number of relevant microposts;
- average number of endorsements of top N microposts;
- average size of influence of top N users.

Additionally, we list the keywords used for the relevance assessment, and provide a link to the list of Twitter identifiers of the relevant microposts.

To illustrate this, an example is shown in Figure 2 of the graphical output of the first proof-of-concept implementation. For this example, the abstract of a WikiNews article[5] was used as input, of which the keywords *Duncan Smith, Iain Duncan Smith, minute-long tirade, axe, holiday weekend, tee shirts, Willie Black, bedroom tax*, and *Capita-sponsored talk* were extracted by AlchemyAPI and used as queries. The result shows a straightforward graph indicating the number of relevant microposts found, as well as the average number of endorsements and user influence size of the top 1000 microposts.

5. DISCUSSION & CHALLENGES

Our first, naive implementation immediately exposed a number of important challenges of the proposed approach, which we discuss in this section.

5.1 Storage and Querying

The first of these challenges is storage. At an average of over 4 million tweets per day, the amount of space needed

[2]http://dev.twitter.com/docs/streaming-apis
[3]http://lucene.apache.org/solr/
[4]http://www.alchemyapi.com/
[5]http://en.wikinews.org/wiki/Thousands_take_to_streets_protesting_%27ratbag%27s_Bedroom_Tax?dpl_id=713879

for data storage and indexes increases quickly. Once a critical size is reached, query speed becomes an issue as well. Queries with single or multiple keywords require a full text search over the data, which can take a lot of time. Therefore, alternative storage and query options are an important aspect to consider for future implementations.

5.2 Natural Language Processing

Apart from these architectural challenges, dealing with social media content is very different from dealing with news articles. Current natural language processing (NLP) tools are tailored to deal with, grammatically correct, well-spelled and full sentence text. Microposts are often very short, are full of slang and abbreviations, and do not contain full sentences. Therefore, standard approaches such as NER fail. Newly developed algorithms such as TweetNLP [6] could potentially offer solution here. Additionally, we need to take multiple languages into account. For example, it should be possible to assess the impact of an English article on Japanese social media. This is one of the less explored but highly relevant challenges in this field.

5.3 Evaluation

Lastly, the question remains of how to evaluate the accuracy of this approach. Due to the highly subjective nature of the concept of social media impact, this is not a trivial task. Furthermore, the approach contains components which need to be evaluated individually as well, such as the keyword extraction, relevant micropost retrieval and mapping to levels of endorsement/influence. Here, crowdsourcing the evaluation using Human Intelligence Tasks, as applied in [11], might be a viable solution. Additionally, the entire application must be evaluated as a tool for newsworthiness assessment by a panel of experts in the field of journalism studies.

6. CONCLUSION

We described a novel approach for assessing the relationship between news and social media. The approach is deliberately kept general, to allow for several possible implementations, each with their own optimizations. We provided a first, naive implementation ourselves, and identified the most imminent challenges for this approach. Finally, we outlined the actions needed to tackle these challenges, and to evaluate the approach properly.

7. ACKNOWLEDGMENTS

The research activities in this paper were funded by Ghent University, iMinds (a research institute founded by the Flemish Government), the Institute for Promotion of Innovation by Science and Technology in Flanders (IWT), the FWO-Flanders, and the European Union.

8. REFERENCES

[1] S. Asur and B. Huberman. Predicting the future with social media. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 492–499. IEEE, 2010.

[2] M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill, and J. Lin. Earlybird: Real-time search at twitter. In *28th International Conference on Data Engineering (ICDE)*, pages 1360–1369. IEEE, 2012.

[3] C. Chen, F. Li, B. C. Ooi, and S. Wu. Ti: an efficient indexing mechanism for real-time search on tweets. In *Proceedings of the 2011 international conference on Management of data*, pages 649–660. ACM, 2011.

[4] P. Dillon Scott. A year of tweets–2012's most popular & important Twitter updates - http://sociable.co/social-media/a-year-of-tweets-2012s-most-popular-and-important-twitter-updates/.

[5] W. Gao, P. Li, and K. Darwish. Joint topic modeling for event summarization across news and social media streams. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1173–1182. ACM, 2012.

[6] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, pages 42–47. Association for Computational Linguistics, 2011.

[7] A. Hermida. From tv to twitter: How ambient news became ambient journalism. *Media/Culture Journal*, 13(2), 2010.

[8] A. Hermida. Twittering the news. *Journalism Practice*, 4(3):297–308, 2010.

[9] L. Jabeur, L. Tamine, and M. Boughanem. Uprising microblogs: A bayesian network retrieval model for tweet search. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 943–948. ACM, 2012.

[10] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web*, pages 591–600. ACM, 2010.

[11] M. Larson, M. Soleymani, M. Eskevich, P. Serdyukov, R. Ordelman, and G. Jones. The community and the crowd: Multimedia benchmark dataset development. *MultiMedia, IEEE*, 19(3):15–23, 2012.

[12] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *4th International Conference on Weblogs and Social Media*, 2010.

[13] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th international conference on World wide web*, pages 621–630. ACM, 2010.

[14] N. Naveed, T. Gottron, J. Kunegis, and A. Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *ACM Web Science Conference 2011*, pages 1–7, 2011.

[15] Z. Papacharissi and M. de Fatima Oliveira. The rhythms of news storytelling on twitter: Coverage of the january 25th egyptian uprising on twitter. In *Presented at the World Association for Public Opinion Research Conference*, volume 312, page 3188, 2011.

[16] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, pages 385–388. ACM, 2009.