Vilnius University Press

Contents lists available at Vilnius University Press

Žurnalistikos tyrimai
2022, 16, pp. 138–151

ISSN 2029-1132 eISSN 2424-6042
DOI: https://doi.org/10.15388/ZT/JR.2022.5

# SmartNews: An Automatic Approach for Event Detection on Media Platforms

**Hussein Hazimeh**
Lecturer, Faculty of Science, Lebanese University, Beirut, Lebanon
Email: Hussein.hazimeh@ul.edu.lb

**Abstract.** *Social Media Platforms (SMPs) are currently the leading media data sources in the world; billions of people's electronic devices have adopted these SMPs for their use. The users ' accounts on these platforms generate massive amounts of data daily. Data have become an essential building block for many organizations of different domains. Recently, media organizations started using social media as a principal source to collect data, mainly news. Having recognized the importance of SMPs and data availability, media organizations are not using these data efficiently. Many media organizations still use and analyze internet data, especially from social media, manually, which leads to many disadvantages. This research proposes a more efficient and automated approach to collecting information from social media. Actually, this paper proposes an integrated framework that can extract data from multiple SMPs and merge them, store them, and finally allow media workers to extract fundamental data (events) automatically and smartly from social media. The proposed framework takes input from a query and finds the following information: top tweets, total likes and retweets on this query, user's identity, sentiment analysis, and finally, the prediction component that can classify if a particular item has classified an event or not. An advantage of this approach is to help media leaders control and track their performance in the media sector and maintain popularity on the internet. The proposed system has been validated on real datasets collected from different data sources. Findings show that this proposed system has remark-*

**Received**: 2022/12/15. **Accepted**: 2023/01/30

*able accuracy, precision, and recall results, after evaluating different machine learning algorithms,*
**Keywords:** *Social data analysis, digital media, social networks, machine learning, data integration.*

## Introduction

Currently, the data in the world have two essential features: Being generated rapidly and being collected from multiple sources. The effect of these features has contributed remarkably to the emergence of the Big data era defining multiple proprietary features (Ghasemaghaei, 2017). Firstly, the key feature of big data is the scale characterized by exponential growth. To handle that growth; new data storage platforms have emerged while other platforms are under development and testing. The second feature of big data today is the lack of structure. In this context, many challenges must be addressed and assessed otherwise; data will not be beneficial for computer systems. The third feature is sparsity since the current existence of technological devices and applications like smartphones, laptops, wireless sensors, network architecture, the internet of things, smart cities, and more real systems generate a large amount of data. Data have no unique source, yet different sources. The last feature of big data is speed, i.e., data are generated rapidly. Therefore, raising the need and necessity for new organizations that work primarily with data to adopt modern adaptation policies to update their technological infrastructure and increase financial investment in technology, among others, to remain successful.

In this paper, we highlight our research on a fundamental Big data source which is social media. According to Power and Phillips-Wren (2011), "Social media (SM) and Social Media Platforms (SMPs) have emerged because of web 2.0 (participatory web) innovations to enhance human communication and create dynamic and interactive dialogues" (p. 251). Alalawneh, Al-Omar, and Alkhatib (2022) add that SMPs are global and are known for their permanent access at any time and anywhere. Thomsen Trampedach (2022) reports that "Around six (6) bil-

lion user accounts and two (2) billion active users can be attributed to only 10 Social Media Platforms (SMPs)" (para 1). These users generate timely content across multiple platforms. The content is diverse such as text, images, and videos. In addition to the common social media users, most organizations in the world create powerful and public profiles on social media to reach their targeted communities (Bashir et al., 2022). Major media companies transformed into social media companies, such as newspapers, TV, shows, and more. In this context, most of the media audience in the world follows the latest news and updates through their social media accounts (Schwaiger, Vogler, and Eisenegger, 2022). Kemp (2022, July 21) posits that "The Reuters Institute for the Study of Journalism (RISJ) published the 2022 edition of its Digital News Report. It asserts that when it comes to news channels, people are now two and a half times as likely to turn to social media for news as they are to turn to physical newspapers and magazines" (para 44). Despite the noteworthiness content of media pages on social media, sometimes users might not be interested in the content published by these pages (Schwaiger et al., 2022). For instance, take a user who follows a page on Facebook that publishes 20 posts daily on different subjects. This user is interested in five (5) subjects only. Such an incident proves that there is no customization. Until now, social media channels do not offer customization features. Emphasizing this fact, Ghasemaghaei (2017) posits that certain users can only unfollow a particular page; they cannot customize the preferred news they would be interested in.

There are many research papers published on finding events in social media (e.g., see Chen, Xu, and Mao, 2019; Halimi and Ayday, 2020; Abousaleh, Cheng, Yu, and Tsao, 2021), but only a few of them focus on the problem of content selection of media events. Based on the above papers, among others, and the scarcity of published resources, in this paper, we address the problem of social media news event detection in an automated and smarter manner. So, the aim is to create a novel system to address the problem. The proposed system deployed several matching and machine-learning algorithms to enhance its accuracy.

The rest of this paper is structured as follows: Section 2 lists several related research papers, followed by section 3 which introduces and explains the proposed system contribution. Section 4 addresses the experimental results and evaluation. Finally, section 5 concludes this work and discusses future directions.

## Literature Review

The recent literature in the context of social media analytics is comprehensive. The focus of this study is to analyze media content on social media platforms. In particular, the related work is divided into two parts: (1) Research on social media content merging and (2) research on social media page popularity prediction.

### *Social media content merging*

Morales, Gionis, & Sozio (2011) proposed a big-data-based solution to investigate and address the problem of social media content matching. They proposed three algorithms and compared them to assess the effectiveness of their matching architecture. Their experiments have been conducted on Flickr and Yahoo! and reached good results in terms of accuracy.

Agichtein, Castillo, & Donato (2008), introduced a framework for detecting high-quality content on social media, mainly Yahoo! They were able to separate high-quality content from spam and fake comments written as user reviews.

On the other hand, some research papers focused on the problem of matching the content of user profiles to link these profiles with each other and merge them. In recent research, in this context, Halimi and Ayday (2020) worked on matching user profiles on social media platforms, specifically Twitter, Foursquare, Google Plus, Twitter, and Flickr. The authors relied mainly on public attributes and match them using a deep learning model, and their results have shown high accuracy compared to others. Other researches have a broader context and cover the

problem of matching entities, where an entity could be a user profile, a page, or a post (Peled, 2013).

Despite the broad research done in the domain of matching content on social media, most of this research has been conducted on user profiles; and there are no up-to-date approaches proposed to match user-generated content on social media platforms.

### *Social media page popularity prediction*

The research published in this domain is notable so far. Many papers have tackled the issue of popularity prediction on social media in different facets, such as images, videos, likes, shares, and more. Chen, Kong, Xu, & Mao (2019) worked on the popularity prediction task leveraging deep learning solutions. They consider two main issues to address, the noisy content of social media posts and the adaptation of deep learning algorithms. Gelli, Uricchio, Bertini, et al. (2015) proposed a system to predict the popularity of social media images by using sentiment analysis and features related to the context of the image. A similar more recent research by Moniz and Torgo (2019) addressed the problem of the popularity prediction of photos on Flickr. The authors utilized multiple deep-learning models to achieve better performance results. In addition to the contributions to this research subject; some surveys have reviewed this domain. Abousaleh, Cheng, Yu, & Tsao (2021) have reviewed dozens of research papers on popularity prediction considering multiple social media platforms. Also, some papers have addressed the problem of event detection in specified languages such as Arabic (Daoud & Daoud, 2020; Rafea & GabAllah, 2018), Chinese (Almerekhi, Hasanain, & Elsayed, 2016; Wang, Guo, & Wang, 2021), and more.

In conclusion, the related work provided in this section is suitable and beneficial for the research community. However, in the context of content matching, the contributions are still minimalistic. Accordingly, the authors of this paper propose a content-based merging approach applied to social media, particularly Facebook, Twitter, and Instagram.

Furthermore, the proposed system contains a prediction component that can predict the popularity of each social media post.

## Materials and Method

### *SmartNews: Proposed System architecture*

In this section, we illustrate our system architecture. Mainly, we divide this architecture into four-ordered components. The first component of SmartNews is to select and describe the media sources, i.e., websites, blogs, and social media platforms. The second step constitutes scrapping the data from these resources to get them all in a single unified dataset. This dataset will be processed and analyzed. After the dataset is obtained we propose some similarity measures applied to the data to capture similar contents. The third step involves a classification task to classify the media content in our dataset. Then, we represent and store our collected events and news inside a knowledge graph. Finally, end users will be notified about the data generated by the system. The agents inside this system architecture represent the media applications or websites used by end-users.

The key objective of this architecture is to extract and classify selected topics to be received by media users. Each of the phases inside the architecture is described in Figure 1.

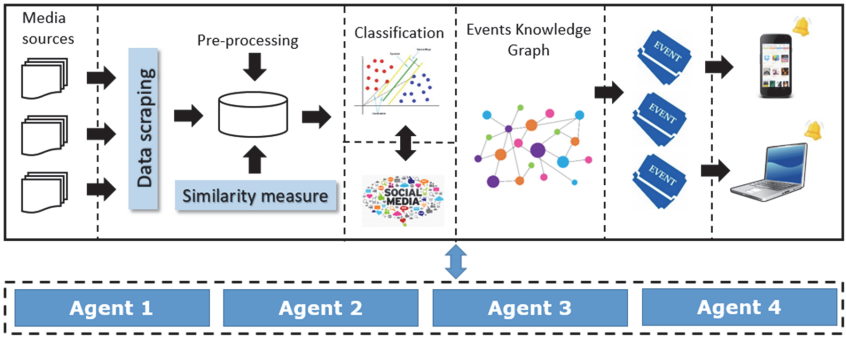


**Figure 1.** Proposed SmartNews architecture

## Process description

### *Media sources*

A media source is any media page on a social media platform that generates regular and consistent content to be received by end users. For instance, The New York Times newspaper. The selection of these sources is dynamic, where we can add or remove sources as needed.

### *Data Scraping*

The data available inside every media source is digital, yet it needs to be extracted in an automated mechanism and a high-performance manner. This objective is realized using a data scraper technique called Selenium (https://www.selenium.dev/). Selenium is an efficient software tool that supports multi-language development intending to scrap content from websites. It provides libraries to enable the scraping of website content smartly and accurately. The main important features to extract from these pages are listed in Table 1.

**Table 1:** The description of all features extracted from every social media page

| Feature | Description |
|---|---|
| Content | The textual content of the item |
| Reactions count | The total number of reactions |
| Comments | The total number of comments |
| Time window | The average delay between all comments on a single item (post) |

The time window is calculated using the following formula:

$$D = \frac{\sum_{i=1}^{n} t(p_{i+1}) - t(p_i)}{n}$$

Where

$n$: Is the total number of comments,

$t(P_{i+1})$ and $t(P_i)$: These are functions that return the time differences

between two comments, and D: Is the average of all time differences between all comments divided by the total number of comments.

After scraping the content using Selenium from the pre-defined media sources, we apply several text pre-processing steps to enhance the quality of the data and raise its performance once analyzed. The main steps followed in the pre-processing are: stop word removal, stemming, and tokenization.

Once the data is cleaned and ready for processing, we store it inside a repository, specifically inside the MySQL database software platform. Then we utilize some similar functions applied to this data as an advanced prep-processing step. For instance, we utilize the Cosine Similarity to compute the similarity between two texts. In this phase, we compare comments that are the same but written by different users. For instance, if two comments are the same but have two different profile names, one of them is deleted, and the other is kept.

### *Classification*

After the data have been prepared and turned into high-quality data. In addition, we have stored such data inside a repository. We prepare our dataset comprising the features described in Table 1 and apply some machine learning supervised algorithms to classify whether the post is an event. The machine learning algorithms utilized in this process are SVM, Decision Tree, Random Forest, and Naïve Bayes.

The dataset used in the classification is composed of four features and the class. The features are shown in Table 1. This classification problem is binary, and the two classes are 'event' and 'not event'.

### *Knowledge Graph*

The extracted events by the machine learning algorithms are stored inside a knowledge graph. A knowledge graph is a knowledge store graph that represents real-world entities inside a graph-based structure. The advantage of this representation is to enhance the semantical aspect

of data and enable high-speed retrieval. For instance, the user (agents in our architecture) can extract customized events from the knowledge graph along with their details.

### *Experiments and evaluation*

In this section, we present the results of our system. In addition, we show the facts about the data used for analysis and results.

The dataset is collected using Selenium (see the architecture). It is composed of 500 instances, and each instance is classified manually as an event or not an event item.

Four main media data sources were involved in the analysis. Although, we can include more or different sources as well without any problems or updates on the system. Table 2 depicts for each source the number of items (posts) that participated in the analysis.

**Table 2:** Data source sizes

| Data source* | Size |
|---|---|
| S1 | 450 |
| S2 | 400 |
| S3 | 430 |
| S4 | 440 |
| *S: digital media source | |

Figure 2 illustrates the data extracted from Facebook as stored and represented inside MongoDB, the repository used to comprise the data used in our analysis.

Figure 3 shows the same data stored in MongoDB converted to semantic relations and stored as a knowledge graph inside the GraphDB repository.

Figure 4 illustrates the scalability results of the proposed system. We were able to maintain a stable speed regardless of the number of posts analyzed increased. The scalability is an important evaluation metric in

our system because the number of posts is arbitrary, and we have to expect a high potential increase in the number of posts, hence designing a highly scalable system is necessary.

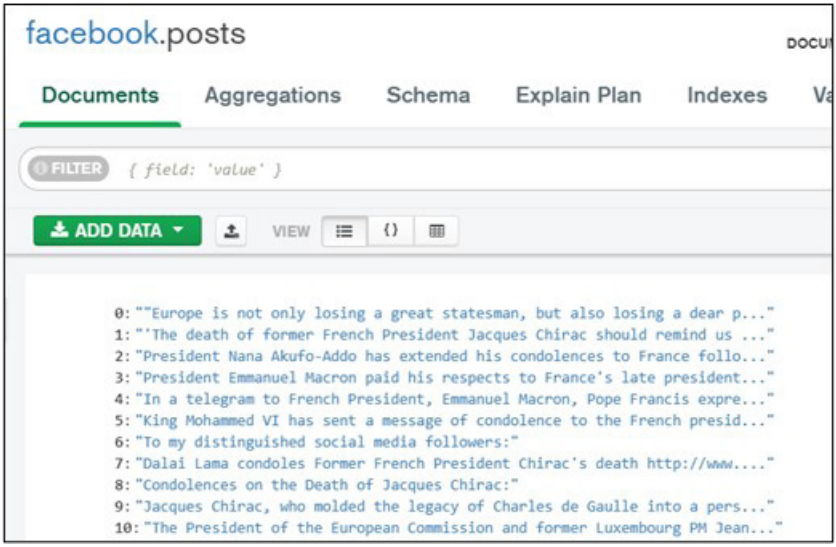


**Figure 2.** The data saved in MongoDB

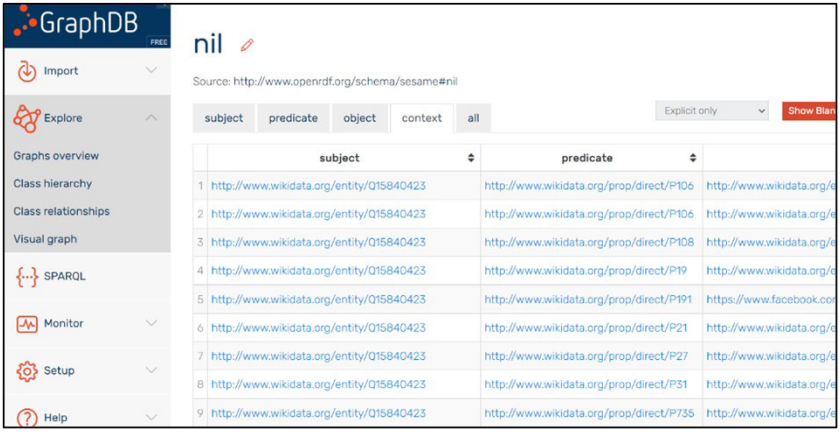


**Figure 3.** The data saved as a knowledge graph triples inside GraphDB

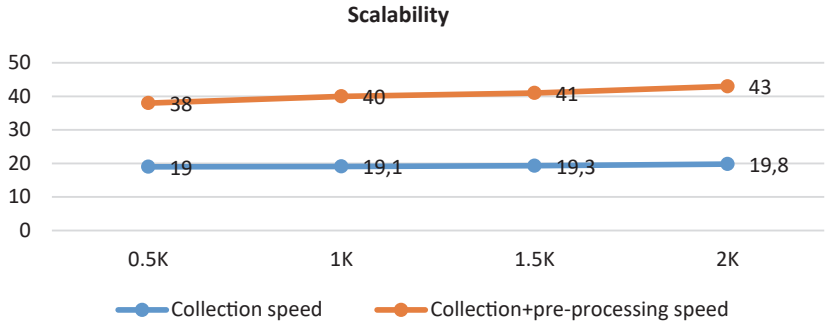


**Figure 4.** The scalability results of the proposed system

**Table 3:** The performance evaluation of pages compared to classifiers

| Evaluation metric | Classification algorithm | Page 1 | Page 2 | Page 3 | Page 4 |
|---|---|---|---|---|---|
| **Accuracy** | SVM | 93 | 94 | 92 | 95 |
| | Decision Tree | 89 | 88 | 91 | 93 |
| | Random Forest | 95 | 94 | 90 | 91 |
| | Naïve Bayes | 93 | 93 | 90 | 91 |
| **Precision** | SVM | 90 | 91 | 91 | 94 |
| | Decision Tree | 93 | 90 | 91 | 93 |
| | Random Forest | 94 | 92 | 95 | 94 |
| | Naïve Bayes | 93 | 90 | 91 | 93 |
| **Recall** | SVM | 93 | 90 | 88 | 94 |
| | Decision Tree | 88 | 91 | 88 | 91 |
| | Random Forest | 94 | 90 | 94 | 90 |
| | Naïve Bayes | 93 | 90 | 93 | 90 |
| **F1-measure** | SVM | 93 | 94 | 92 | 95 |
| | Decision Tree | 92 | 95 | 92 | 95 |
| | Random Forest | 91 | 93 | 91 | 93 |
| | Naïve Bayes | 90 | 91 | 90 | 91 |

The results presented in Table 3 show the comparison of the four classification algorithms' performance on four social media pages. In particular, we used accuracy, precision, recall, and f-measure.

**Table 4:** The average score of each evaluation metric

| Evaluation metric | Classification algorithm | Average score |
|---|---|---|
| Accuracy | SVM | 92.4 |
| Precision | Decision Tree | 91.2 |
| Recall | Random Forest | 92.5 |
| f-measure | Naïve classifier | 91.3 |

Based on the data presented in Table 4, we conclude that all classification algorithms have roughly the same performance. However, SVM has the highest performance in terms of accuracy compared to other algorithms. However, the Decision Tree has the highest precision, and the Random Forest has the highest recall compared to other algorithms. Finally, the Naïve classifier has the highest f-measure compared to the other algorithms. Table 4 shows the average score per each metric applied to each algorithm.

## Conclusion

To benchmark the evaluation results presented in Table 3, we conduct our experiments on three events: covid-19, protests, and currency changes that happened in 2021 in Lebanon. After we get the results from our system, we compare its results with the real events that we have selected.

In conclusion, our proposed system can extract and classify events from news in high-score evaluation results. The results of this system validate that it can be utilized by any media organization to find important events on social media pages.

In this research paper, we illustrated the importance of social media data's role in facilitating and increasing the performance of media companies. Specifically, we address the problem of social media event detection from multiple social media platforms. Our system searches for interesting content on these platforms and retrieves the highest important ones (events). To classify the event from non-event content,

we proposed a machine learning-based solution that performs perfectly on this task. In addition to the news retrieval component, our system is composed of a prediction component that can predict the event from non-event of the retrieved posts.

Large-scale media organizations can operate the proposed system in this research. These organizations can automate the task of news/post search and retrieval. Therefore, reducing the effort and time spent on these tasks while doing them manually. Moreover, it can increase the financial outcome by decreasing the capacity of human resources required to do these tasks.

## References

Abousaleh, F., Cheng, WH., Yu, NH. and Tsao, Y. (2021) 'Multimodal Deep Learning Framework for Image Popularity Prediction on Social Media,' *IEEE Trans.* Cogn. *Dev. Syst.* vol. 13, no. 3, pp. 679-692

Anisa, H., Ayday, E. (2020) 'Profile Matching Across Online Social Networks. Information and Communications Security,' *Proceedings of the 22nd International Conference, ICICS 2020,* Copenhagen, Denmark, August 24-26, 2020, pp. 54-70

Agichtein, E., Castillo, C., Donato, D., Gionis, A. and Mishne, G. (2008) 'Finding high-quality content in social media,' *WSDM*, pp. 183–194

Alalawneh, A.A., Al-Omar, S.Y.S. and Alkhatib, S. (2022) 'The Complexity of Interaction between Social Media Platforms and Organizational Performance,' *J. Open Innov. Technol. Mark. Complex.*, vol. 8, p169. https://doi.org/10.3390/joitmc8040169

Almerekhi, H., Hasanain, M. and Elsayed, T. (2016) 'EveTAR: A New Test Collection for Event Detection in Arabic Tweets,' SIGIR 2016: 689-692

Bashir, E., Hejase, H.J., Danash, K., Fayyad-Kazan, H. and Hejase, A.J. (2022) 'An Assessment of Students' Preferences Using Social Media Platforms on Their Selection of Private Universities in Lebanon,' *Journal of Business Theory and Practice*, 10(3), pp1-39. https://doi.org/10.22158/jbtp.v10n3p1; URL: http://dx.doi.org/10.22158/jbtp.v10n3p1

Chen, G., Kong, Q., Xu, N. and Mao, W. (2019) 'A neural popularity prediction model for social media content,' *Neurocomputing*, 333, pp. 221-230.

Cui, Y. and Wei, Y. (2021) 'Chinese Text Event Detection Technology Based on Improved Neural Network,' FSDM, 2021, pp. 436-442.

Daoud, M. and Daoud, D. (2020) 'Sentimental event detection from Arabic tweets,' Int. J. Bus. Intell. *Data Min.* vol. 17, no. 4, pp. 471-492

Gelli, F., Uricchio, T., Bertini, M., Del Bimbo, A. and Chang, SF. (2015) 'Image Popularity Prediction in Social Media Using Sentiment and Context Features,' *ACM multimedia,* pp. 907-910.

Ghasemaghaei, M. (2017) 'The Effects of Operational and Cognitive Compatibilities on the Big Data Analytics Usage: Firm Distinctive Value Creation,' *AMCIS* 2017

Kemp, S. (2022, July 21) '*Digital 2022: July Global Statshot Report,*' Data Portal, [Online]. Available at: https://datareportal.com/reports/digital-2022-july-global-statshot (Accessed: 22 November 2022).

Moniz, N. and Torgo, L. (2019) 'A review on web content popularity prediction. Issues and open challenges,' *Online Soc. Networks Media,* vol. 12, pp. 1-20

Morales, G.D-F., Gionis, A. and Sozio, M. (2011) 'Social Content Matching in MapReduce,' Proc. *VLDB Endow*, vol. 4, no. 7, pp. 460-469.

Peled, O., Fire, M., Rokach, L. and Elovici, Y. (2013) 'Entity Matching in Online Social Networks. International Conference on Social Computing,' *SocialCom 2013*, SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013, Washington, DC, USA, 8-14 September, pp. 339-344

Power, D.J. and Phillips-Wren, G. (2011) 'Impact of Social Media and Web 2.0 on Decision-Making,' *J. Decis. Syst*., 20, pp249–261. https://doi.org/10.3166/jds.20.249-261

Rafea, A. and GabAllah, N. (2018) 'Topic Detection Approaches in Identifying Topics and Events from Arabic Corpora'. *ACLING 2018,* 270-277

Schwaiger, L., Vogler, D. and Eisenegger, M. (2022) 'Change in News Access, Change in Expectations? How Young Social Media Users in Switzerland Evaluate the Functions and Quality of News,' *The International Journal of Press/Politics*, vol. 27, no. 3, pp. 609–628. https://doi.org/10.1177/19401612211072787

Thomsen Trampedach. (2022) '*Social Media Monitoring'*, [Online]. Available at: https://www.thomsentrampedach.com/online-brand-protection/social-media-monitoring/ (Accessed 21 November 2022).

Wang, Z., Guo, Y. and Wang, J (2021) 'Empower Chinese event detection with improved atrous convolution neural networks,' *Neural Comput. Appl.*, vol. 33, no. 11, pp. 5805-5820.