Mediaeval 2021 Emerging News: Detection of Emerging News from Live News Stream Based on Categorization of News Annotations

Omar Meriwani[1],
[1]Scientific Editor, Real Sciences website and magazine
omar.meriwani@gmail.com, omar@real-sciences.com

ABSTRACT

This paper describes the contribution of RS_OMERIWANI in the Mediaeval 2021 Emerging News task. Among the various definitions of emerging news, this work is based on the definition of emerging news as the type of news that would gain more attention from news sources, i.e. higher frequency in publishing the same news. Relying on the categorization of the news annotations, the classification process has been completed through an unsupervised clustering to generate training data for a supervised neural network model that classifies the news based on the categories that are mentioned in it. The accuracy score for the final model was 74%, with a 65% F-Score for detecting emerging news. The final model fulfilled the requirements of newsworthiness and completeness of reported events as well as the relevance criteria in the task evaluation.

1 INTRODUCTION

Journalism is in an ongoing challenge of detecting news angles [1] that both interest the readers and satisfy the objectivity requirements of news. Giving this mission to the computer, requires specifying the exact meaning of emerging news, such as the different concepts discussed in [2]. This paper describes the contribution of RS_OMERIWANI in the Mediaeval 2021 Emerging News task [3].

Having sufficient news samples, it is possible to have the computer perform the first part of this task, allowing for automation of the selection process of determining emerging news. News usually gets unbalanced interest from news sources; some news stories get published in more than ten main sources of news within a specific country/region/language, while some other news never gets the same level of attention, being published only in one or two sources.

This work is done using supervised and unsupervised machine learning models and by relying on categories to find the news that has higher chances of getting published more frequently in the media.

2 APPROACH AND METHODS

Our approach focuses on what the *more frequently published news* could be, and, on news annotations provided from the News Hunter platform [4] which can provide both the news live stream as well as the fine categorization of named entities that are mentioned in the news.

In Table 1, we can see some of the news samples with the number of times they got published during the same two-hour time window. Some gain a lot of attention while others never get published by more than one or two news sources.

Table 1: News pieces with the number of times they got published on different websites on 22nd October 2021

News	Frequency
Haitian gang leader threatens to kill kidnapped missionaries	26
Last Known Photos Of Brian Laundrie & Gabby Petito Together	11
UK palace says queen, 95, spent night in hospital for checks	14
Braun Strowman Says WWE Turned Him Into A Corporate Monster	1
Record number of daily vaccinations in Dominican Republic	1
EXCLUSIVE: Vicki Gunvalson Addresses Breakup With Steve Lodget	2

This approach is based on the preference of media channels, regardless of any deep analysis of the news content. We assume that the attention that some news articles may get is based on the nature of named entities it deals with, for example, the first sample in Table 1 contains the following categories:

a gang leader, an island, and members of a religious group (missionaries).

It seemed more interesting than the fourth sample which has categories combination that includes:

American, wrestler, World Wrestling Entertainment

To achieve this approach, the work is divided into two parts:

1- **Finding similar news**: Using a technique to cluster the similar news together, we created new labels based on the clustering results, and labelled the final training dataset with 1 or 0 based on the threshold of appearing three times or more. For example, the first new rows in Table 1 would be labeled as (1) and the latter three rows

MediaEval'21, December 13-15 2021, Online

would be labeled as 0. The label (1) indicates the emerging news.

2- **Supervised classification:** Using the resulting dataset, we have vectorized the categories of the news annotations and used the vector representation as a training data for an artificial neural network to predict the labels mentioned in the previous step, either (1) for emerging news or (0) for other news.

2.1 Unsupervised news clustering

News titles were transformed using a term frequency–inverse document frequency (TFIDF) vectorizer [5] in order to create vectors that could be used in the clustering algorithm.

K-Means algorithm was used to make N/2 clusters of the original news dataset with total N samples. In that way, it's assumed that the number of clusters will be no less than half the total number of samples, enabling us to detect the similarity of the news more accurately.

Due to the computational complexity of the clustering using a high number of clusters, the original dataset of ~50K news samples was divided into 7 batches.

The final dataset included 12,667 news samples. 4,879 of them had been published three times or more by different news sources, and 7,788 news samples had only been published once or twice. The clusters were used to create labels that indicate clusters, and the news within the same cluster were labeled as (1) or (0) according to the cluster size, the data was also balanced for labels (0) and (1).

2.2 Supervised classification

After the news dataset was created mainly by the indicators provided by K-means clustering, the results were ready for supervised learning model that can classify the data using different set of features, namely, the categories of news annotations. The News Hunter platform already provides annotations for the main named entities and a set of classes for these annotations. We used a count vectorizer for each news' categories set.

We used a multiple-layers perceptron with hidden layers sizes of (20,20,20). The data was divided into 2:8 for testing and training.

The output format was then structured by returning the news titles of emerging news as well as the keywords that were extracted using the Single Rank algorithm [6] which is implemented in the Kex Python package.

3 RESULTS

Table 2: Frequency of Special Characters

Label	Precision	Recall	F-Score
0	0.76	0.86	0.8
1	0.72	0.58	0.65

The accuracy achieved by the model was 74%, while the F-score for emerging news detection was 65%. Precision for emerging news – label (1) – is shown in Table 2, which is close to the precision of label (0). Based on the recall score, it could be said that the model may flag many varieties of annotations' categories as false negatives.

The independent human evaluation results stated about whether the newsworthiness and completeness requirements were satisfied: "Yes, the information provided brings insights that can conform an event and provides extra information with the keywords that can help to get a fast overview of the reported story. The keywords add some extra information which is not present in the title, helping the journalist to better understand the event.". The evaluation also described the relevance aspect: "provides potentially relevant events for journalist or not widely covered events that may have not been yet seen by journalists"

4 DISCUSSION

Categories of named entities or annotations in the news could be used as an auxiliary feature to support more comprehensive models. However, the results of categories alone have could fail to detect some emerging news by flagging them as false negatives. Some aspects would still not be covered, such as the sentiments of verbs that indicate violence which may usually get more attention.

The final output that could be extracted using this method also lacks some features that would make it rich enough to be more realistic for journalists.

However, the model would still be efficient for working with poor news details; it can work regardless of the text length, links availability, or titles format, because it is only based on the categories of the main entities. Logically it can work on many cases when it deals with the abstract attributes of the named entities that are mentioned in the news.

REFERENCES

[1] Marc Gallofré Ocaña and Andreas Lothe Opdahl. 2020. Challenges and opportunities for journalistic knowledge platforms. Proceedings of the CIKM 2020 Workshops. Galway, Ireland.

[2] Marc Gallofré Ocaña, Lars Nyre, Andreas Lothe Opdahl, Bjørnar Tessem, Christoph Trattner, Csaba Veres. 2018. Towards a big data platform for news angles. The 4th Norwegian Big Data Symposium (NOBIDS).

[3] Marc Gallofré Ocaña, Andreas L. Opdahl and Duc-Tien Dang-Nguyen. 2021. Emerging News task: Detecting emerging events from social media and news feeds. MediaEval'21: Multimedia Evaluation Workshop.

[4] Arne Berven, Ole A. Christensen, Sindre Moldeklev, Andreas L. Opdahl, and Kjetil J. Villanger. 2018. News Hunter: building and mining knowledge graphs for newsroom systems. NOKOBIT.

[5] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," *Carnegie-mellon univ pittsburgh pa dept of computer science,* 1996.

[6] C. G. Broyden, "The convergence of single-rank quasi-Newton methods," *Mathematics of Computation,* no. 24.110 , pp. 365-382, 1970.