# Building Event Graphs from GDELT Streams

Media Futures

*Marius Alexander Pedersen, Master's in Information Science,*
*Department of Information Science and Media Studies, UiB*
*Marius.A.Pedersen@student.uib.no*

## Abstract

**GDELT  is an enormous dataset of events and information extracted from news articles across the globe every 15min. I aim to extracted these events and use machine learning to cluster together events that are connected to better represent the events on a higher-level event structure.**

## Research question

**Can the low-level events reported by the GDELT project be used to represent news events as knowledge graphs?**

The events from GDELT are on a low level in that they depict only where, what category of event, and in some cases who interacts. There is therefor a hope when we cluster together events that may be sub events of an event we can better depict higher level concepts and maybe an evolution of how the event unfolded.

I already have run an agglomerative clustering on a extract from the dataset and seen some good signs in what events it clustered together. I have begun applying word embeddings on the information available from the dataset to better calculate each cluster of events. So that events like a landslide will be closer to an event like earthquake, than an event of a prime minister visiting another country.

In evaluating the correctness and usefulness of the finished product I aim to compare what is stored and represented from my program vs what big industry news aggregators store and relates to the same events.

Example if there is an event registered that is for both lets say Google news as an example and mine. Then compare them with what does Google News represent as correlated and what information are they presenting on the topic vs what mine have as related and what information is presented. This is a way of comparing with what has been accepted from big industry as acceptable and will for this serve as a "gold standard" to compare against


## PARTNERS

## HOST

UNIVERSITY OF BERGEN

## FUNDER

*This research is funded by SFI MediaFutures partners and the Research Council of Norway (grant number 309339).*