**Data Story**

**Do we care about all?**

Image Description (Number of Articles vs Number of Events):

*The number of events reported in the period 2015-2017 are grouped by the countries where they happened and the number of articles sums up all the GDELT news articles these events received. Typically, an event is mentioned in about 5 to 10 news articles, but this number can exceed 100 and more. The size of the bubbles is proportional to the population of the country.*

**How large is our circle of empathy? (maybe need to change some of the titles/ do an adaptation between this and the following titles)**

**Do emotions define whether an event reaches a certain level of importance?**

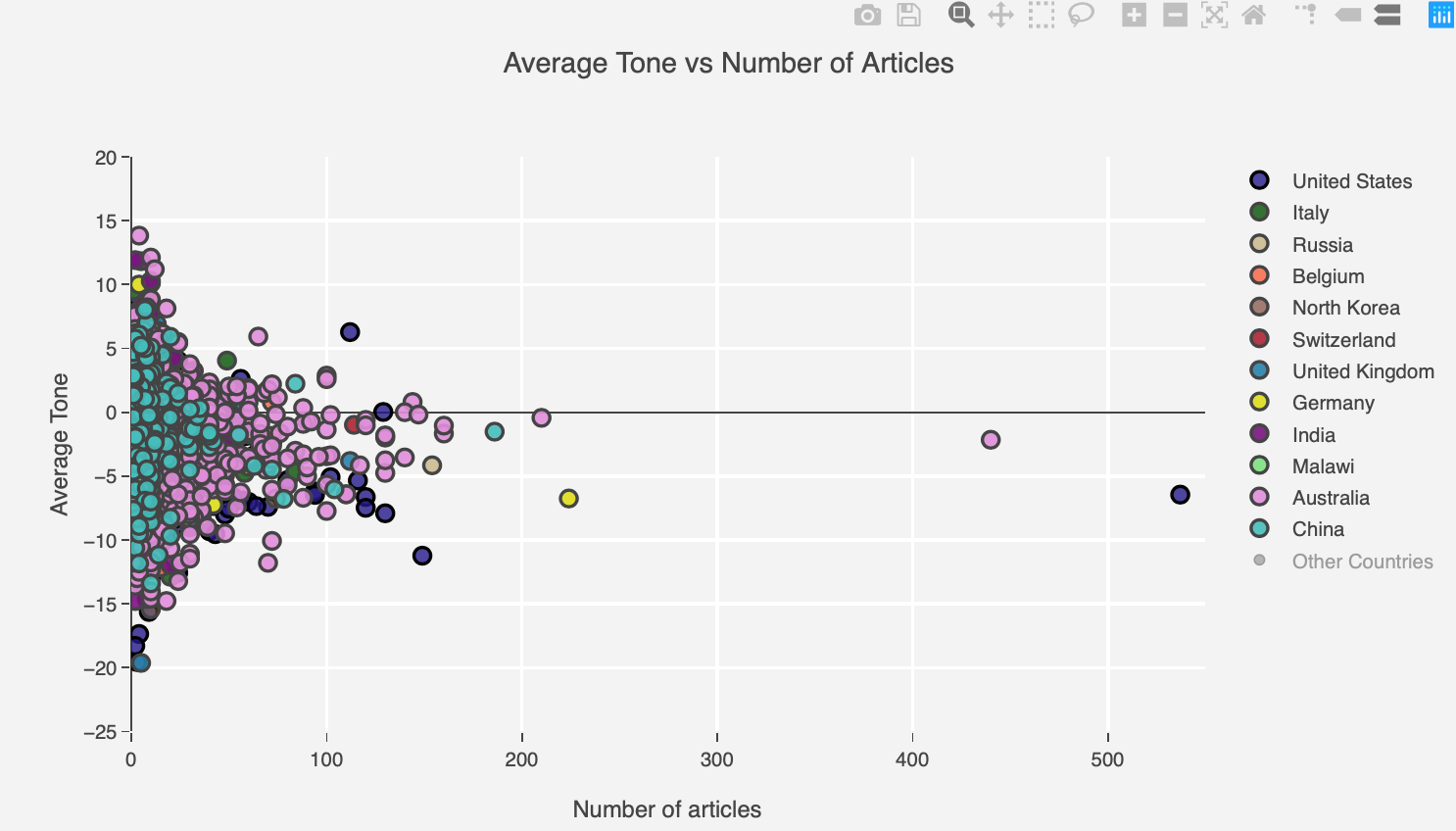
We expect that high-impact events are associated with a larger spectrum of emotions, and that events such as an unexpected terror attack which affect a large number of people and which are spread rapidly in the news could be frequently seen in the GDELT dataset. However, as seen in the figure XX, events tend to be labelled as emotionally neutral when reaching a certain level of importance. On the other hand, minor events reported by only a few sources are having a much larger variation in the reporting tone. While this might sound surprising at first, there might be a simple underlying explanation: averaging the document tone over all kind of news articles, from neutral news agency reports, to more emotional boulevard articles, results in an average tone which tends to lose the distinct emotions of the individual sources. Nonetheless, the figure reveals that in general the news tend to more negative than positive, and events with a significant number of mentions in the news are rather found on the negative side of the distribution.

Image Description:

*The average tone of the articles reporting an event within the first 15 minutes it is first seen and the number of articles reporting the event within this short period. The countries indicate where the event happened. Only a fraction of the all the GDELT events in the 2015-2017 period is plotted (0.01%).*

**What about the distance between the event and the news article source?**

If it is not the importance of an event that shapes the emotional tone of an article, is there maybe another feature? From the previous analysis, we gathered that it is rather difficult to associate emotions to events, and that the overall positiveness and negativity of an event might be washed out if calculated by an aggregation of several news sources. We thus dig a step further, and focus on the individual news article reporting an event and we try to see whether the distance between an event and the news article source influence how we perceive the incident. We define an emotional charge which integrates the polarity and the tone of an individual article and look for a trend when plotted against the geographic distance of event and source location.

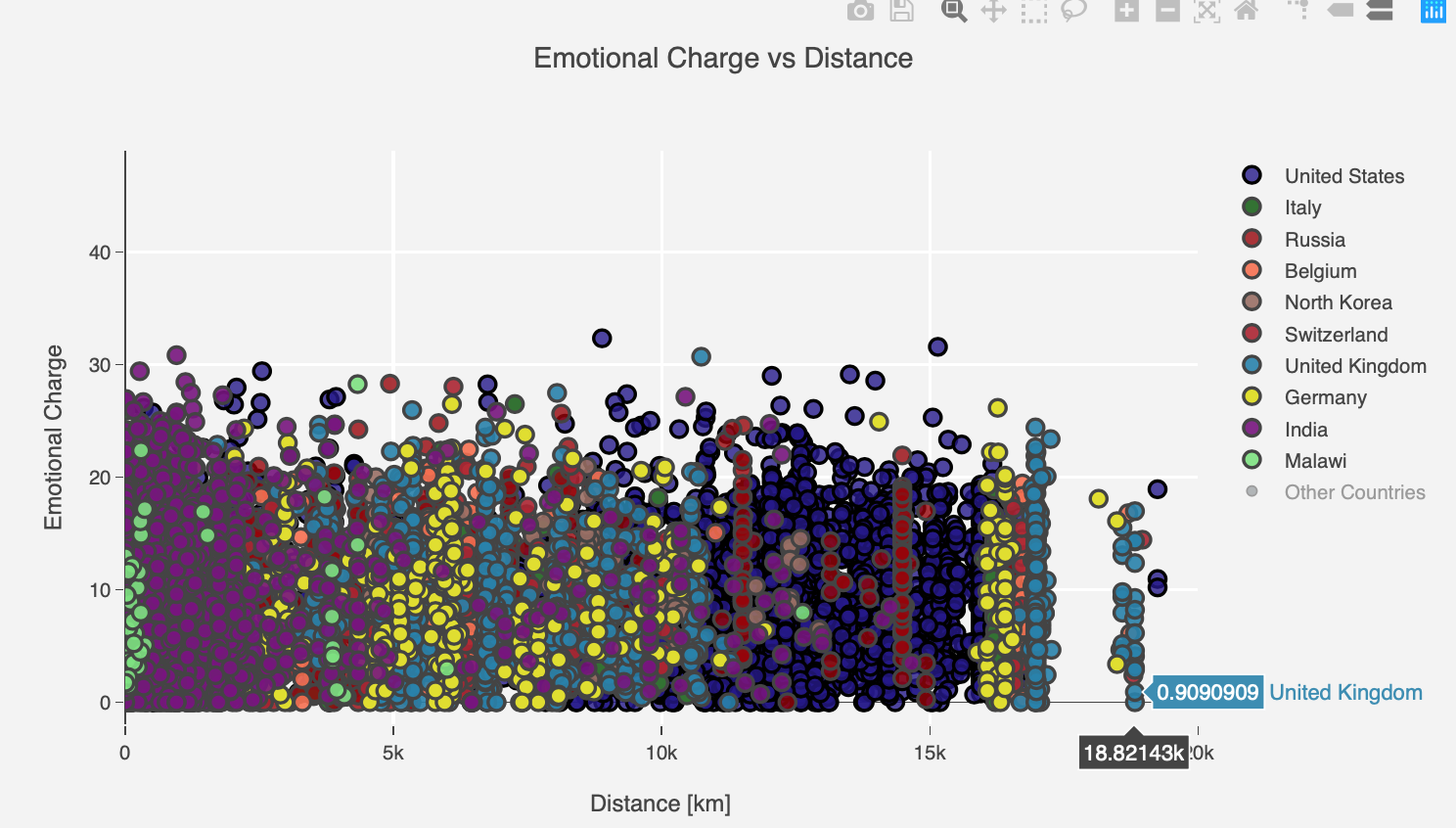


Image Description:

*The emotional charge of an article (measure of polarity and tone) vs the geographic distance between an event and the news source calculated by the Great-Circle distance formula. Only a fraction of the all the GDELT events in the 2015-2017 period is plotted (0.01%).*

The figure speaks for itself: the number of kilometres separating the event and the source is not in relation with the emotional charge. What about the instinctive impression that controversial events accompanied with emotional debates reach a larger fraction of the population? If we have a closer look at the figure, we see that events occurring in countries where generally a lot of events occur and which are rather countries playing an important role in the international politics have a higher chance to be mentioned all around the globe, than events happening in countries less known in the world (as example, we chose Malawi, a small African country). Hence, it seems as if a breaking news, independent of how emotional or not it is, can reach a large number of people.

**Are we emotionally predictable?**

The complexity behind what actually defines the emotions of a news article paper is obviously a function of the words that are used to describe the event. Words are abundant, and thousands of words are at the disposition of the news reporter. Attributing a tone to the document is done by GDELT through the count of positive minus negative words, which implies that the resulting tone is not necessarily the tone a human reader would attribute to the document. In addition, a neutral tone, meaning a zero score, could be the consequence of equally abundant negative and positive words (this issue is however mitigated by providing a polarity score: a high polarity score, but a zero average tone would point towards such a situation).

Analysing the semantics of a document through the words is computationally expensive. We try to overcome the computational hurdle by predicting the emotions of a document by a simpler collection of features, such as the location of the event, who reported it, what kind of event was it, where there religious concerns, … which additionally allows to see whether the emotional reaction of the news can already be predicted by just knowing basic facts of an event.

Figure of avgtone distribution and features

Applying a random forest machine learning algorithm and splitting on a training and test, we get an accuracy of around .. %. While this accuracy is very low, the model does indeed give an indication of the emotional tone of an article and outperforms a random and a uniform model (a model randomly predicting one of the average tones, and a model only predicting the category “negative” which is the most prevalent). However, the model is clearly not satisfying and the question of how accurate the emotional tone actually is remains. Naturally one would say, that events where someone got killed are more negative, but in the dataset such events also get attributed to positive emotions. On the other hand, events reporting the release of a hostages, have negative scores which looks surprising. Let’s look at some examples which all have the attribute “KILL” to understand this situation:

Description:

*News headline taken from the GDELT dataset. Indicated is the measured tone of the article and its polarity in parentheses. Explanation of the polarity?*

The first news headline undoubtedly merits the strong negative score, and the second which describes a situation where a kid could have possibly died, but survived thanks to medical intervention, merits the positive score. The difference between this first and last example can however only be captured by differences of the choice of words, where the last document indeed uses a more positive vocabulary than the first document.

The last example, however, results from a misclassification and shows that the word analysis is not a hundred percent reliable and representative of the actual document emotion, and underlines that simple words, without the context, define the tone given by the algorithm.

To keep things simple, and without including all the words such as the GDELT algorithm, we integrate a small collection of positive-minded words in our model. The result is striking: by only adding positive words, the accuracy could be improved by … %.

Figure of positive words:

This analysis highlights that only knowing basic facts of an event can provide a good indication of how it will be perceived in the media in the different countries. However, since the emotional metrics provided by GDELT only relies on word counts and thus introduces a lot of noise into the data, a precise model cannot be derived.

**Things to be mentioned in the conclusion (somewhere at the end of our analysis):**

First 15 minutes update might be a reason for significant biases towards countries at the forefront of reporting digital news such as the UK with BBC and the US with ..?