*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4343–4351
Marseille, 11–16 May 2020


# The Connection between the Text and Images of News Articles: New Insights for Multimedia Analysis

**Nelleke Oostdijk**[1], **Hans van Halteren**[1], **Erkan Başar**[2], **Martha Larson**[1]

[1]CLS, Radboud University, Nijmegen
[2]FloodTags, The Hague
{n.oostdijk, hvh, m.larson}@let.ru.nl, basar@floodtags.com

**Abstract**

We report on a case study of text and images that reveals the inadequacy of simplistic assumptions about their connection and interplay. The context of our work is a larger effort to create automatic systems that can extract event information from online news articles about flooding disasters. We carry out a manual analysis of 1,000 articles containing a keyword related to flooding. The analysis reveals that the articles in our data set cluster into seven categories related to different topical aspects of flooding, and that the images accompanying the articles cluster into five categories related to the content they depict. The results demonstrate that flood-related news articles do not consistently report on a single, currently unfolding flooding event and we should also not assume that a flood-related image will directly relate to a flooding-event described in the corresponding article. In particular, spatiotemporal distance is important. We validate the manual analysis with an automatic classifier demonstrating the technical feasibility of multimedia analysis approaches that admit more realistic relationships between text and images. In sum, our case study confirms that closer attention to the connection between text and images has the potential to improve the collection of multimodal information from news articles.

**Keywords:** Corpus (Creation, Annotation, etc.), Multimedia Document Processing, Tools, Systems, Applications

## 1. Introduction

Online news is a natural source of information about current events. News articles consist of components including headlines, article text, images, and image captions. In order to build a system that can automatically extract event information from online news, it is important to understand the connection between these components. In this paper, we report the results of a case study that we carried out in order to gain a better understanding of this connection.

Our larger goal, which is the wider context for this paper, is to create a system that can automatically collect and aggregate information about the impact of flooding from online sources. Such information supports disaster managers in monitoring and responding to flooding events as they evolve, and also helps scientists gain a better understanding of flooding by allowing them to analyze historical patterns (FloodTags, 2019). Event information from online sources provides an important complement to information that can be derived from remote sensing imagery. Specifically, it is not dependent on factors such as the periodicity of the satellite or cloud-cover conditions. Further, it reflects not only the development of the disaster, but also the specific impact that the disaster is having on people. Our aim is to fully take advantages of both the text and the image information in online sources, and for this reason we are focused on developing multimedia analysis techniques.

The motivation for our case study derives from issues that we encountered in our larger mission to develop multimedia analysis technology for disaster management. We previously produced a data set to study flood information in social media (Bischke et al., 2018). With this data set we noticed that there was not a straightforward relationship between images and text, and we decided to turn to online news. We produced another data set to study flood information in online news sources (Bischke et al., 2019), but again we had underestimated the complexity of the connection between text and images. A brief survey of the literature, discussed in Section 2, revealed that the field of multimedia analysis at large makes radically simplifying assumptions concerning how text and images are related. We recognized that a case study directed at gaining more insight would be helpful to our own mission. At the same time, we realized that the results of such a study would also serve as much-needed evidence for the multimedia analysis community about how their underlying assumptions must evolve in order to more effectively tackle the problem of extracting information from data that consists of a combination of text and images.

With this work, we release an annotated data set with links to 1,000 online news articles (Oostdijk et al. Flood News Multimedia Analysis Data, 2019), and we also make the following contributions: First, we point out, using a concrete case study based on manual analysis, the inadequacy of the assumption that an image accompanying a text is described by that text. Currently, this assumption is widely adopted in the field of multimedia analysis. We discuss how the relationship between image and text is different for headlines, captions, and the body of the article. Second, we separate news articles by topical aspects, each corresponding to a different message that the article is intending to communicate. Not all topical types involve a one-to-one relationship between a news article, and a single, ongoing event. We show that different topical types of article use different types of images. Third, we validate the manual analysis with experiments using an automatic classifier, which confirm that multimedia analysis techniques have the potential to be sensitive to factors impacting the connection between image and text.

The paper is organized as follows. In the next section, we provide a brief overview of the related work on multimedia analysis and on language resources for studying flooding. Then, Section 3 provides detail on the general use scenario which provides the context for our case study, and also on the data we use. Section 4 forms the heart of the paper. It describes the methodology of our manual analysis and details our results. Section 5 is dedicated to the automatic analysis which we carry out as a validation of the manual analysis. Finally, in Section 6, we conclude and provide an outlook.

## 2. Background and Related Work

The field of multimedia analysis develops techniques capable of extracting information from data that is composed of different modalities. Multimedia research assumes that a message is conveyed by multiple media, and that if only one medium is analyzed, part of the message is missed (Friedland & Jain, 2014). However, multimedia research often makes highly rigid assumptions about the relationship among the media. Specifically, in research that deals with combinations of text and images, the text is usually assumed to describe the image. Other possible connections between text and images, for example, specification, complementarity or emotional enhancement do not receive adequate research attention.

A multimedia research area that illustrates this narrow focus on description is automatic image captioning (Bernardi et al. 2016; Bai & An, 2018; Hossain et al., 2019). This area develops algorithms that generate natural language sentences describing the content of an image. Typically, the descriptions represent what a generic person would report as visible in the image. Expert opinions on what is visible in the image are not taken into account. In short, image captioning assumes that the connection between text and image is that the text provides a simple, literal description of an image. Little automatic captioning work diverges from this narrow characterization. A notable exception is Bitten et al. (2019), which takes image context into account. Sharma et al. (2018) release a data set of images accompanied by captions for the purpose of developing automatic captioning systems. The data set is created by scraping images and the alt-text descriptions from the web. The authors filter out all images for which none of the text tokens in the caption can be mapped to the visual content of the image.

Note that we do not claim that all multimedia analysis work adopts an overly simplistic conceptualization of how text and images relate. However, we find the lack of research on realistic connections between text and images is serious enough that it may hold back the state of the art in multimedia analysis for disaster management. This concern motivated us to carry out the case study we report in this paper.

Our work is currently of particular relevance since journalists have started to actively adapt the ways in which they cover climate related disaster news. An example is the recently initiated effort by the British newspaper *The Guardian* (Shields, 2019). This effort is following advice by the outreach project "Climate Visuals", which includes recommendations to use less stereotyped images for depicting the image of natural disasters (Climate Visuals, 2019). Systems that automatically collect information about climate related disasters such as flooding need to be aware of how image and text connect, and also need to be able to keep up as these connections continue to evolve.

## 3. Use Scenario and Data Resources

### 3.1 Use scenario

The context of this work, as mentioned in Section 1, is a wider project creating a system that can automatically collect and aggregate information from online news articles on flooding. We are particularly interested in extracting a timeline of flooding events that can be used directly by disaster managers to determine what they have missed in their analysis of remote sensing imagery. The timeline should contain information on the time and the place of the event, but also about the magnitude of the impact, e.g., number of casualties, number of buildings destroyed.

While pursuing this aim, we found that automatic methods were challenged by news articles providing information on flooding that goes beyond description of isolated incidents, for example, comparisons of currently ongoing events to past events, or analysis of seasonal trends. When we collected our data and designed our analysis, we took into account that future work might be interested in the broader spectrum of flood-related information in the news beyond the individual events. For this reason, we designed our data set so that our case study would focus on events, but also give us insight into other types of information.

We chose to focus on flooding in Southeast Asia (particularly Myanmar and Cambodia) since that is currently of major concern to disaster managers, and thus also to FloodTags. Southeast Asia has a surprising number of English-language online newspapers. We focus on these newspapers since we do not have a command of Southeast Asian languages, and also to allow the results of our case study to be fully accessible to the English-speaking scientific community.

### 3.2 Data collection

Our case study is carried out on a selection of 1,000 news articles extracted from English-language news sources in Myanmar and Cambodia.

First, we identified a wider set of English-language news sources in Southeast Asia using online resources such a Wikipedia and also search engines. We included the English language versions of newspapers published in, e.g., Khmer and Burmese. From these we collected news articles containing at least one flood-related keyword. Keywords used included *flood*, *floods*, *flooding*, *flooded*, *inundation*, *inundations* and *inundated*. All articles supposedly are about flooding (by water). Note that the appearance of a keyword does not guarantee that the article will be related to flooding by water, since the lemmas *flood* and *inundate* can be used in many senses. As will be seen, our manual analysis investigates the line between *flood* used in the context of a flooding event, and *flood* used in other senses. The difference is not a sharp as one might expect.

Next we made a selection of the 1,000 articles that we would study. The selection is not completely random, but we took measures to ensure that it is representative enough for the purposes of our study. We focused on a subset of the major English-language news sources from Myanmar and Cambodia. Since we are interested in the connection between text and images, we used only news articles that included an image. Since we are particularly interested in captions, we excluded cases in which an image had no caption. If there were multiple images included with a news article, we considered only the first image. Most images were photographs; occasionally there were satellite images or maps. These were retained in the collection. The 1,000-article collection contained a handful of duplicates, i.e. the same article appears twice in the collection (in different newspapers or in the same newspaper on two consecutive days). Occasionally the same paragraph appears in more than one article. Statistics describing the 1,000-article collection are provided in Table 1.

| source | # articles | publ. year |
|---|---|---|
| bnionline.net (news) | 58 | 2015-2019 |
| elevenmyanmar.com (news) | 56 | 2018-2019 |
| mizzima.com (news, domestic, regional) | 161 | 2015-2019 |
| mmtimes.com (national news) | 371 | 2010-2019 |
| phnompenhpost.com (national) | 354 | 2001-2019 |

Table 1: Composition of the 1,000-article data set of online news articles used for our case study (by source).

All articles were published between March 2001 and August 2019. We note that the majority of articles were published after 1 January 2014; only 285 appeared before that time, of which 56 before 2008. This is consistent with our desire to focus on the most current journalistic practices. Articles were rather varied in length (33-55,041 words; mean 680, sd 1777).

# 4. Manual Analysis

## 4.1 Objectives and methodology

The design of our use case study was inspired by the following line of questioning: If humans read articles and look at images, they find that they cover multiple aspects of flooding, and that the image is not always related in the same way to the text. Can we find enough systematicity in the topical types of news articles on flooding to propose a categorization? How is this systematicity reflected in the article headline? Can we find patterns of co-occurrences between topical types of text, and images depicting different contents? How are these patterns reflected in the image captions?

We are interested in human views of text because online newspapers are designed to be read by humans. We adopt the standpoint that understanding the relationship of text and image from a human point of view must precede exploration of the ability of automatic approaches to differentiate between different topical types of news articles and different types of images. For this reason, the manual analysis described in this section forms the major contribution of our study, and the automatic analysis in the next section is intended only as a verification and a demonstration of future potential.

We adopted a manual content analysis methodology that is adequate for the exploratory nature of our use case study. One expert human annotator (first author) is fluent in English and has extensive experience in the area of text analysis and computational linguistics. To perform the manual analysis, she sequentially studied all 1,000 articles in the data set and coded them with topical categories. When encountering a news article that did not fit into an existing category she created a new category. The whole article was considered, but the article text was used as the main source of information for the decision. No annotation protocol was defined in advance, but rather it emerged as the annotator worked her way through the sequence of articles. As mentioned above, although the focus of our wider work is on events, here, we are interested in all the topical aspects represented by news articles, and these were taken into account in the categorization. The advantage of having the manual analysis of the entire collection by a single person is the improved self-consistency of the resulting categories. A potential disadvantage could be that it might be difficult to reproduce them; therefore we investigated (Section 5) whether an automatic analysis could achieve a similar classification.

Simultaneously with the coding of the articles with topical categories, the annotator investigated a range of other aspects important for the relationship of image and text including what was shown in the image, whether the headline already hinted at the article class, and whether the caption of the image contained any information besides image name and type, and source, which would place the text and the image in the same time and geographical space.

## 4.2 Results: Text

### 4.2.1 Topical categorization

The manual analysis process resulted in six dominant categories representing different topical aspects of flooding, plus one category for articles that did not fit within these dominant categories. The topical categories and their relative distribution are presented in Table 2.

| | |
|---|---|
| TF<br>**F**lood, ongoing<br>119 articles | The article is about an ongoing flooding event in a specific area at a specific time; there is mention of houses and roads being flooded, people and cattle that have drowned, schools being closed etc. |
| TA<br>**A**ftermath of flood<br>166 articles | The article focuses on the aftermath of (recent) flooding event; it describes the situation as the water is receding, the effects in terms landslides that have occurred, people that were displaced, the damage sustained, relief activities ongoing in the area and such. |
| TT<br>**T**hreat of flooding<br>42 articles | The focus of the article is on the immediate threat of flooding; the article typically mentions rising water levels, (expected) heavy rainfall, etc. |
| TL<br>**L**ong-term follow-up<br>27 articles | The article describes the events following flooding events that took place (possibly) weeks to months earlier: this can be about authorities or politicians talking about preventive measures, the setting up of support programmes, etc. |
| TG<br>**G**eneral phenomenon of flooding<br>30 articles | In the article the focus is on (the threat of) flooding as an example of what could happen, for example, as a result of a cyclone passing in the area, an earthquake, and such. Other than articles in class TT, flooding here is more of a phenomenon than an event. |
| TP<br>**P**re-flooding measures<br>34 articles | The article is about preparing for possible flooding, taking measures to prevent future flooding. |
| TX<br>583 articles | None of the above. See text for detailed description. |

Table 2: Topical types: Our manual analysis revealed clusters of articles associate with six dominant topical aspects of flooding, which we assigned category labels.

The TX category contains two major subsets of articles. In the first subset (which we refer to as TX-flood), are articles that refer to flooding without flooding being the main focus of the article. The articles in this category are quite diverse: they are about elections, drainage infrastructure, dam construction, deforestation, diseases, fishery, education etc. Articles such as those on deforestation and drainage infrastructure are flood-related as they identify main (potential) causes of flooding. In other articles, flooding is typically described as part of the circumstances that (may) exist at some point; examples here are articles about elections and school exams taking place during flooding, or about protests against the construction of a dam because this presents a threat to the habitats of wildlife.

In the second subset of TX (which we refer to as TX-non-flood, the keyword in the article has a different sense than flooding by water. Articles in this class cover a wide range of topics, including for example the Rohingya crisis, armed violence, immigration, tourism, economics, and politics, in which we encounter flooding by people (refugees, tourists, addicts), products (imports, drugs, fake medicines, timber, chemicals, firearms, cars, …), lies, help, messages, and inquiries. This second subset contains a substantial number of cases (116).

While trying to establish the focus of the news articles the annotator was forced to make a decision in order to identify a single focus for the article as a whole. Sometimes this decision was quite difficult. This problem arose not just with the lengthier news articles; even with relatively short articles it appeared that one part of the article would, for example, focus on ongoing flooding and another on the aftermath or the threat of flooding. In such cases typically what is reported on is what is happening in a larger region: some part was flooded earlier and there the water is receding, another part is actually experiencing flooding and yet another part is under threat of being flooded soon.

In sum, we have found that it is possible to identify dominant categories related to different aspects of flooding. We observe that there is a surprisingly large amount of news coverage of flooding that does not focus on a specific flooding event, and also that articles that do focus on events may do so partially. Next we move on beyond considering the text of the articles to look more at the role of the headlines.

#### 4.2.2 Headline

We are interested in headlines since it would be beneficial to multimedia analysis approaches if the headlines could be used in place of the text in order to determine the topical category of a flood-related news article. As described above, the data were collected on the basis of keywords. In 259 articles an instance of a keyword was found to occur in the article headline; most of the articles were about a current flooding event (91 articles) or about the aftermath (105 articles). Typical headlines are for example *Flood persists in Ayeyarwady Delta* or *Residents fear landslides as Hpakant floods*.

There were a few cases where the focus of the article was not on flooding, yet the headline contained one of the keywords. Examples are *Emotional farewell for flood volunteer* and *Flood-struck high school reopens.* For articles where the keyword was used in a different sense (article subclass TX-non-flood) we came across headlines such as *Chinese longyis flood Mandalay market* and *Bad medicine floods countryside.* We conclude that the headline has a potentially useful contribution to make to determining the topical category of a news article.

### 4.3 Results: Text and Image

Next, we build on the results from the textual analysis to look at what our manual analysis revealed about the connection between images and text.

#### 4.3.1 Images

At the same time that the articles were coded for topical category, the annotator developed a set of codes reflecting the depicted content of the images. For these codes, the image content was leading, i.e. the annotator primarily looked at what could be seen in the image and ignored the information in the caption. The image categories and their relative distribution is presented in Table 3.

| | |
|---|---|
| IF<br>**F**looding<br>254 articles | In the image actual flooding can be seen, i.e. in photographs water can be seen in places where it should not be, while in maps areas are indicated that according to the legend are flooded. |
| ID<br>**D**amage from flooding<br>55 articles | The image shows damage to buildings, roads, bridges, train tracks and other infrastructure, trees, statues, mine shafts etc. Other than in images in the IF class, in none of the ID images water can be seen and from the image alone it is impossible to tell whether the damage was caused by flooding or must be attributed to some other cause (e.g. an accident, fire, earthquake or cyclone, or vandalism). |
| IL<br>**L**andslide aftermath<br>5 articles | The image shows a landslide. From the image we cannot tell what caused the landslide: it may be flooding, heavy rain but also, for example, an earthquake. |
| IP<br>**P**erson or people<br>200 articles | The image is a portrait of a person or it prominently shows one or more people. Typically they are government officials, politicians, or other people that are well-known (generally or more locally). |
| IX<br>486 articles | None of the above. See text for detailed description. |

Table 3: Image types: Our manual analysis revealed five clusters of images which we assigned content-related category labels.

The "IP" category of images depicting people was surprisingly large. We observed that especially with articles that take a human-interest angle, there are images that portray the person who features in the news article. Strikingly, in each case the caption with the image identified the person by name and often also their title or function (e.g., *Chinese Foreign Minister Mr Wang Yi*, *Myanmar's foreign minister Aung San Suu Kyi* or *AMDP candidate Nai Win Htut*, but also *U Aung Myat, a native of Min village in Mandalay Region*).

The "IX" category of images that did not fit in the other four categories also deserves additional comment. These images are quite diverse, ranging from images showing (anonymous) people to landscapes, wildlife, and construction sites but also boats at a water festival. Here we

should note that where people appear in the image, the focus is not on the specific people but rather on the scene as whole: what the image is showing is, for example, flood victims who receive goods from a relief organization, workers filling sandbags, or protesters at a demonstration. In the caption the references to the people that appear in the photographs are all indefinite noun phrases (e.g. *a resident of Pauk village, a woman, tourists, children*).
We also note that in some cases it was not easy to tell what exactly the image was showing. Typical problematic cases were the images showing rice paddies or floating houses where it was unclear whether there was flooding or not.

#### 4.3.2 Text-Image connections

Next, we turn to the question of which categories of images occur with which categories of articles. Table 4 presents a co-occurrence table containing the raw counts of text and image categories.

| Text/Image category | **IF** Flood | **ID** Dam-age | **IL** Land -slide | **IP** Peo-ple | **IX** |
|---|---|---|---|---|---|
| **TF:** Flooding | 96 | 3 | 0 | 4 | 16 |
| **TA:** Aftermath | 75 | 19 | 2 | 16 | 54 |
| **TT:** Threat | 24 | 1 | 0 | 0 | 17 |
| **TL:** Long-term | 11 | 3 | 2 | 4 | 7 |
| **TG:** General | 5 | 3 | 0 | 9 | 13 |
| **TP:**Pre-flood | 12 | 3 | 0 | 2 | 17 |
| **TX:** Flood | 29 | 21 | 1 | 121 | 295 |
| **TX:** Non-flood | 2 | 2 | 0 | 44 | 68 |

Table 4: Raw co-occurrence numbers: Content categories of images and topical categories of articles

Our first observation is that articles about current flooding (TF) tend to include an image showing flooding (IF), whereas with articles focusing on the aftermath of flooding (TA) there is also quite a number of images showing damages (ID), people (IP, e.g., ministers and other officials visiting the disaster struck area) and other photographs showing, for example, flood victims. Next, we observe that there is a relatively large number of articles focusing on the threat of flooding (TT), which include a flooding image (IF). A possible explanation is that these would be photographs of flooding events in the past, used to recall what that was like. While this is true for some images, it does not explain all cases. This observation will be discussed further in Section 4.3.4 "Temporal distance" below.
We also observe photographs of flooding (IF) with articles that focus on something else than flooding. When we take a closer look at, for example, the images found with articles in class TX we find that here IP and IX images are preferred. In the case of an IP image, the function of the image is, presumably, to show what the person(s) mentioned in the article looks like. Images in the IX class form a miscellaneous category in the sense that if an image could not be classified as IF, ID, IL or IP then automatically it ends up in this class. This explains why the images are diverse. The articles in which the keyword is used in a different sense (TX-non-flood) are mostly associated with images of types IX (68) and IP (44), but also with two of type ID and even two of type IF. If we follow the early literature on Word Sense Disambiguation (Gale et al., 1992), we consider it to be surprising that a word should be used in more than one sense in one and the same news article. However, after investigations on SemCor and DSO (Krovetz, 1998) such cases are no longer entirely unexpected.

#### 4.3.3 Image captions

Now, we turn to report our observations on the distribution of image captions. As noted above, the 1,000-article set that we analyze was selected such that all articles have an image with a caption. We note that most captions in our dataset are made up of a single sentence or phrase (1-76 words, mean 20.6, sd 12.6). Not all captions, however, are rich in content, i.e., 40 captions only include an image reference (name and type, e.g., *flooding.jpeg*) or the image source (typically *Photo by* [name]). 'Information-rich' captions overall tend to briefly summarize what the image is showing and relate the image to the article text by identifying the people in the image or the event. A fair number of captions (277 out of the 960 rich-content captions) contain a mention of one of the flooding-related keywords used to create the data set. In all, 134 of these occur with images showing flooding (IF), the majority (110) with articles about current flooding or the aftermath of recent flooding where the keyword is also found in the headline of the article. Thus there are 62 articles about current flooding with a flooding image and 48 articles about the aftermath of a recent flooding, where both the headline and the caption contain one of the keywords.
Many captions also include a time reference in the form of a date or a temporal expression (*yesterday*, *six months after her home was washed away in the collapse of the dam*) and/or a geo location (from very specific, e.g. *Min Pyar township of Rakhine State, Myanmar* to rather general, e.g. *Myanmar*). In sum, our analysis reveals that the image caption to some extent describes the content of the image. However, in many cases the caption provides additional information which is not conveyed by the image alone.

#### 4.3.4 Temporal and spatial distance

Our analysis provides insight on the role that temporal and spatial distance plays in the connection between the image and the text of a news article. We define temporal and spatial distance as the degree to which the subject matter of the image and the event described by the article took place at the same time and in the same geographical space.
To investigate temporal distance, we calculated the time difference between image and text on the basis of the publication date of the article and the time reference in the caption. Investigating our data, we distinguished between the following four major classes:

1. There is a time lapse between the image and the text of an article of anything from 0 days up to 1 month. [347 cases].
2. The time lapse between image and text is from 1 up to 6 months [71 cases].
3. The time lapse between image and text is more than 6 months [88 cases].
4. No time lapse could be established as the caption did not provide any time reference [482 cases; these include 20 cases where the image is a file photo].

There were 12 cases of news articles that could not be placed in one of the four classes. These included images

with captions for which additional knowledge would be needed in order to be able to infer the date (e.g., images with a caption like *Four days after being ousted from the ruling USDP's top party position, ...)*. They also included images for which the time reference in the caption appears not to be related to the time at which the image was taken (e.g. *The National Election Committee has said that the July 29 polls will proceed as scheduled* …).

The tendency towards more recent images (if we may indeed assume that the time reference in the caption is a reliable source as regards the creation of the image) is not surprising. Since the news articles in our collection are news reports, rather than editorials or feature articles, recency of what is reported in the text and depicted in the accompanying images is to be expected. However, there is also a fair proportion where the image is less recent. We dove more deeply into investigating why this is the case.

Our further exploration started with the conjecture that the temporal distance is correlated with the article category on the one hand and also the image class on the other hand. We already saw that with articles about current flooding (TF; 119 cases) most of the images are flooding images (IF; 96 cases). Of these images 51 are of a recent date while for 43 no date could be identified. We hypothesize that captions which do not include a time reference, are either recent (the image is directly linked to the focus of the article, e.g., describing a current flooding event and the recency is implicit) or timeless, i.e. in the context of the article the exact date of when the image was produced is irrelevant. This observation suggests that where the article is reporting a current flooding event the image is actually about the same event. There are two articles that include an older photograph. One article describes the flooding of refugee camps as a result of the first monsoon rains of this year and the image is one that was taken in one of the camps last year when the camp also suffered flooding. The other article includes a 5 year old photograph showing people in a boat paddling along a flooded street. The article text describes the current flooding situation but stresses the fact that there is no cause for concern as people were well-prepared.

In 39 out of a total of 166 articles on the aftermath of flooding (TA), the image is recent and depicts flooding. The same is true for the 6 images showing damage and the 2 images showing a landslide, 8 showing people, and 18 other images. Some 18 images are less recent, while 74 go without date.

We already noted above that articles that focus on the (immediate) threat of flooding in the majority of cases include an image of flooding. The fact that a flooding image is used can be understood, as it is difficult to imagine how to visualize the threat of an event that has not yet taken place. A natural expectation might be to find (older) images of past flooding events. Surprisingly, however, 12 out of 24 images are quite recent. On closer inspection the explanation for this is rather simple: while the focus of the article is the flood threat, the article also covers flooding that has already struck another place or region and the image actually relates to that event.

A different picture emerges when we look at the images used with articles focused on pre-flooding measures (TP). Here only 4 of the images are recent, 13 are more than one month old and 17 go without date. When we look more specifically at the images showing flooding, 7 of the 12 images date back more than one month (5 even more than 6 months) while the remaining 5 images go without date.

Finally, we turn to discussion of the temporal distance in the case of TX, the category of articles that do not belong to one of the dominant flooding topic categories. Here, we see that 133 images are recent (class 1: between 0 days and 1 month old), 34 are relatively recent (class 2: between 1 and 6 months old), 40 are old (class 3: over 6 months). We distinguished two classes of TX news articles associated with images depicting people (IP). The first involves recent images, and seems to be centered on what person looks like now, or where the person has been. Recency is particularly important when the person is shown at some official event (conference, site visit) since it is relevant to confirm the person's presence at a location at a specific time. The second class involves older images, and is centered on cases where the article describes for example a recent event or development leading to change and the image shows the situation as it existed before.

Up until this point, we have investigated temporal distance. However, we have an important point to make about spatial distance that emerged from our manual analysis. We observed that captions frequently contain information referring to a geographical location. Such information is important for placing the article text and the image in the same geographical space. However, we quickly realized that analyzing the spatial relationship of captions and images requires a detailed knowledge of Southeast Asian geography. It was not possible to simply match the location names in the article and the caption and be certain that they shared the same spatial location. For this reason, we were unable to do a deeper analysis of spatial distance, although it clearly plays an important role in the connection of image and text in news articles.

In sum, our case study revealed that temporal and spatial distance is relevant for understanding the connection between images and text. There is not a single set of conventions for the difference in time between when an image is taken and when the events described in an article occur. This must be taken into account by multimedia analysis methods that attempt to extract information on events by merging information in images and information in text. If the two are not temporally aligned, merging of the two modalities will not be straightforward. The spatial dimension is even more problematic. It is non-trivial to determine the degree to which an image represents a location represented in an image. If data sets are to be created in order to combine location information in images and news article text, it is essential that the ground truth is created by annotators with detailed understanding of the region. The information in the captions does not make it clear to an outsider how distant the place represented in an image is from the place discussed in the newspaper article.

## 5. Automatic analysis

We tested the validity of the categories that emerged from the manual analysis by attempting to reproduce them using an automatic classifier. Success at that task could ultimately also form the basis of an automatic annotation in the future. However, here we limit ourselves to a pilot, which provides a basic demonstration that the categories of news articles and images can be automatically distinguished.

#### 5.1.1 Experimental setup

Given the pilot status of the current experiments, we did not opt for extensive comparative selection of classifiers or for finetuning. Rather we make sensible basic choices in the automatic procedure. For this reason, our results should be taken as a proof of concept and not as an estimate of the annotation quality that can be reached automatically. We expect that in the future substantial improvements will be possible.

We use a random forest classifier (Breiman 2001) in order to carry out multi-class classification. We classify news articles into the seven topical categories, and also into the five image categories. The features of our classifier are cosine similarities between word embeddings representing field instances and word embeddings representing field classes. Note that using similarities between embeddings, rather than using embeddings directly, serves to reduce the size of our feature set. Next we describe the feature extraction in detail.

Our features are formed on the basis of three fields: the full text of the article, the headline and the caption of the image. We first tokenize all fields (leaving casing intact) and build an embedding for each field instance. We chose to use word embeddings due to the limited size of the data set. Vectors of word tokens would quite possibly provide insufficient overlap in non-zero components to calculate cosine distances. Also, because of the limited data, we use pretrained embeddings. In this initial phase of the work we chose FastText vectors (Mikolov et al., 2018), specifically we use 300 dimensional vectors based on 600Bw Common Crawl (Common Crawl FastText, 2019). We built field embeddings by taking a weighted average of the component token embeddings, weights being derived from the IDF of the token and the position in the field (decreasing over the first 200 words and remaining level after that). The IDF is calculated based on about 4000 written texts from the British National Corpus (BNC-Consortium et al. 2007).

In addition to the embedding for individual fields, we built field class embeddings for the seven text classes and the five image classes, by averaging all embeddings for field instances falling in the class.

We then took cosines between each field embedding and each class embedding. Where the field in question belonged to the class in question, we excluded the field embedding from the average, in this way using a leave-one-out strategy of testing. We then normalized the cosines into Z-scores with regard to all cosines for the field embedding in question. For the text topic category prediction we used the similarities of full text field instance to full text field class, headline field instance to headline field class and headline field instance to full text field class, using only the article class models. For the image category prediction, we used the similarities of caption field instance to full text field class and caption field instance to caption field class, including both article and image class models. With these similarities we hope to capture the main focus of the article and the main focus of the caption (which is in turn related to the image).

We use the Weka (Frank et al., 2016) implementation of the random forest classifier. We adopt the default settings apart from number of features per tree. This number is set according to second author's general experience in text classification to 2/3 of the total number of features. For completeness, we mention a detail of the classification pipeline. With an eye to future work, we implemented separate classifiers for sub-classes of TX that were represented by a substantial number of examples (TX dam (40), TX politics (39), TX agriculture (32 articles) and TX forest/wetlands (27)). However, for evaluation they were merged back into TX, as it is the main classes we are interested in here.

#### 5.1.2 Results: Classification of article topics

The confusion table of the automatic text-based classification of news articles into topical categories is shown in Table 5.

| Article category | **TF** | **TA** | **TT** | **TL** | **TG** | **TP** | **TX** |
|---|---|---|---|---|---|---|---|
| **TF** | 77 | 24 | 4 | - | - | - | 15 |
| **TA** | 26 | 106 | - | - | - | 1 | 33 |
| **TT** | 10 | 3 | 20 | - | - | - | 9 |
| **TL** | 1 | 8 | - | - | - | - | 18 |
| **TG** | - | - | - | - | 11 | - | 18 |
| **TP** | 4 | 7 | - | - | 1 | 3 | 19 |
| **TX** | 8 | 23 | 4 | - | 5 | 3 | 539 |

Table 5: Confusion matrix: article classes. Rows are manual annotations, columns are predictions.

Overall, the results indicate a basic ability of an automatic classifier to distinguish the topic categories that emerged from our manual analysis. Of the 116 articles in TX-non-flood (articles with a different sense of the keyword), only 3 are mis-predicted, as TA. Although there is a trend towards correct recognition of the classes, especially for the three larger classes, the current classification is far from perfect. The fact that we chose to use a basic pipeline, as previously mentioned, can explain part of this. The numerous misclassifications on smaller topical categories, notably TL, point towards the inability of the learner to deal with unbalanced data. This is confirmed by the greediness of the TX model.

We can also conclude that the distinction between types TF (Flooding, ongoing) and TA (Aftermath of flooding) is not clear. It would be interesting in future work to reinvestigate these categories to check if they are indeed distinct, or if they should be considered to form a continuum. Note that if we take TF and TA together as a single category, the automatic classification is reasonably able to distinguish this category. Recognizing TF+TA within the whole data set has a precision of 81.5% and a recall of 78.5%, leading to an F-value of 80.0%. In sum, the manual annotation appears to be valid. More data, for all topical categories, would be needed in order to draw more definitive conclusions about the potential of automatic methods to differentiate all topical categories.

#### 5.1.3 Results: Classification of images

The confusion table classification of images in news articles into image categories is shown in Table 6. Here it should be noted that this classification is purely text-based using the text of the caption. We also applied an image classifier for comparison. The results of the comparison are discussed below.

| Image category | **IF** | **ID** | **IL** | **IP** | **IO** |
|---|---|---|---|---|---|
| **IF** | 132 | - | - | 2 | 12 |
| **ID** | 11 | - | - | - | 2 |
| **IL** | 1 | - | - | - | - |
| **IP** | 2 | - | - | 12 | 3 |
| **IO** | 27 | - | - | 2 | 27 |

Table 6: Confusion matrix: image categories. Rows are manual annotations, columns are predictions.

For the automatic classification into image types we decided to process only those 233 articles for which the manual and the automatic annotation agreed that the articles focused on an ongoing flood (type TF or TA). Again the models for the smaller classes, ID and IL, do not manage to claim any articles. For IF, IP and IO, however, classification quality is acceptable. It should be noted that some captions (9 within the 233 articles under investigation here) did not actually have any (rich) content.

Since the larger context of this work is disaster management, we are interested in identifying images of actual flooding. For this reason, we are keenly interested in the ability of the classifier to distinguish images of flooding (IF). With regard to this class, our classifier achieved a precision of 90.4% and a recall of 76.3%, for an F of 82.8%. For comparison purposes, we checked the output of our image classifier, which used the caption, with the output of a computer vision system (Bischke, 2019). This system carries out image analysis and outputs labels representing the content that is visually depicted in images. This manages to recognize flood relatedness with a precision of 93.8% and a recall of 83.6% (F 88.4%). Combination with AND or OR can be used to improve, respectively, precision or recall of both (AND P95.7% R75.3% F84.3%; OR P76.6% R98.6% F86.2%), but always at the cost of a lower F. Improvements in the caption-based recognition and smarter combination methods should prove useful here.

## 6. Conclusion and Outlook

We have reported the results of a case study of 1,000 online news articles related to flooding that investigated the connection between the text of the articles and the images accompanying the articles.

Our general conclusion is that flood-related news articles do not consistently report on a single, currently unfolding flooding event. Further, articles about ongoing flooding often are associated with flood-related images. However, it is not advisable to assume that a flood-related image will directly relate to a flooding-event described in the corresponding article. We have seen that the connection between the text of an article and its image can be a loose one: with large temporal distances and difficult-to-determine spatial distances between the two.

The observations made during our studies point to four roles for images that appearing with news articles:

- Images visualize what the text describes: image of flooded streets with text describing flooding, image of damage with text describing damage inflicted, image of protesters with text describing protest against dam construction
- Images visualize people referred to in the text (identification: what does this person look like), possibly on a specific occasion when this is what the text focuses on (prime minister speaking at a conference)
- Images visualize a situation as it existed before while text describes/suggests how a similar situation may arise (flood threat)
- Images visualize a situation as it exists now but which will be affected by developments described in the text (image of elephant in area that will be flooded once dam is ready)

In short, our use case has firmly established that it is important to abandon the assumption that images depict events described in text.

Additionally, we point out that the concept of an 'event' is also difficult. When does one flood end and another begin? Is flood season an event? When is the flood ongoing? In order to use the results of this study to improve the usefulness of information extracted from online media to disaster managers, it is critical to define in advance a clear use scenario which can be used to establish what should and should not be considered an event.

Future work will be directed at addressing the limitations of this study. We would like to understand how the keyword set used to collect the data has impacted the collection. We would like to carry out the manual analysis with multiple passes and multiple annotators in order to understand the stability of our categories, and get a better estimation of their relative distribution. Finally, we would like to investigate whether the newspapers we studied from Southeast Asia are representative of worldwide journalistic practices, or whether these differ across countries and across the conventions that are adopted by journalists and the conditions under which they practice (relatively well or less-well funded).

Future work in multimedia analysis should not assume that only a single type of connection is possible between text and image. Instead, leveraging the multimodal information contained in online news articles requires a realistic understanding of the connection between text and images.

## 7. Acknowledgements

Thank you to FloodTags and the FloodNews project team, including Simon Brugman and Benjamin Bischke for their help with the interface used for the manual analysis and also with the visual processing of the images.

## 8. Bibliographical References

Bai, S. & An, S. (2018.) A survey on automatic caption generation. Neurocomputer 31, pp. 291-304.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E. Ikizler-Cinbis, N., Keller, F., Muscat, A., & Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. Journal of Artificial Intelligence Research, 55:409–442.

Bischke, B. (2019). Vision Impulse Flood News Analysis System.

Biten, A.F., Gomez, L., Rusiñol, M., & Karatzas, D. (2019). Good News, Everyone! Context driven entity-aware captioning for news images. Proceedings of the Conference on Computer Vision and Pattern Recognition.

Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32.

Climate Visuals. (2019). https://climatevisuals.org/ (accessed 2 December 2019)
FloodTags. (2019). https://www.floodtags.com/ (accessed 2 December 2019)
Frank, E., Hall, M.A., & Witten, I.H. (2016). The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufman, fourth edition.
Friedland, G. & Jain, R. (2014). Multimedia Computing. Cambridge University Press.
Gale, W., Church, K.W., & Yarowsky, D. (1992). One Sense Per Discourse. In DARPA Speech and Natural Language Workshop.
Hossain, MD.Z. Sohel, F., Shiratuddin, M.F., & Laga, H. (2019). A Comprehensive Survey of Deep Learning for Image Captioning. ACM Comput. Surv. 51, 6, Article 118 (February 2019), 36 pages
Krovetz, R. (1998). More than One Sense Per Discourse. NEC Princeton NJ Labs Research Memorandom.
Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)
Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
Shields, F. (2019). Why we're rethinking the images we use for our climate journalism. The Guardian. 18 October 2019.

## 9. Language Resource References

Bischke, B., Helber, P., Zhao, Z., de Bruijn, J., & Borth, D. (2018). The Multimedia Satellite Task at MediaEval 2018. Working Notes Proceedings of the MediaEval 2018 Workshop.
Bischke, B., Helber, P, Brugman, S., Zhao, Z., & Basar, E. (2019). The Multimedia Satellite Task at MediaEval 2019. Working Notes Proceedings of the MediaEval 2019 Workshop.
BNC-Consortium. (2007). The British National Corpus, version 3 (BNC XML edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium, 5(65):6.
Common Crawl FastText, English Word Vectors. https://fasttext.cc/docs/en/english-vectors.html (accessed 2 December 2019)
Oostdijk, N., Halteren, H. van, Başar, E., & Larson, M. (2019). Oostdijk et al. Flood News Multimedia Analysis Data. https://github.com/ErkanBasar/LREC2020-News-Multimedia-Analysis-Data/blob/master/Oostdijk-et-al-Flood-News-Multimedia-Analysis-Data-2019.csv