World conflicts' impact

Nuno M. Mota Gonçalves
nuno.motagoncalves@epfl.ch

Matteo Y. Feo
matteo.feo@epfl.ch

Lucía Montero Sanchis
lucia.monterosanchis@epfl.ch

Abstract

In this project we focus on certain conflicts that would be expected to arise public concern and obtain a numerical approximation of the sentimental impact that said conflicts have on Twitter. We then use the estimated impact for conflicts in several countries to try and find if the impact depends on the country of the conflict. We used the Uppsala Conflict Data Program (UCDP) dataset to identify comparable conflicts in different countries and continents. The main purpose of this project is to shed some light on world-wide situations to which the general public might be oblivious to.

1 Introduction

For our project we decided to make use of both UCDP and Twitter datasets. From these, we would like to figure out any existing gaps in the information proliferation around the globe, regarding the conflicts' locations.

As to how we will make that information clear, we decided to create a predictive model. This model takes a conflicts' category and country as features and tries to predict the sentimental impact on Twitter. To do this, we will use techniques like Named Entity Recognition and Sentiment Analysis to find out which Tweets are worth considering, for a given conflict, and their sentiment strength (not if they are positive, negative or neutral but their strength from 0.0 to 1.0). Analysing the sentiment strength before and after a conflict's start date, we would then define sentimental impact in the same range of values (from 0.0 to 1.0). Applying a threshold to the results would give us either impactful or not impactful - which is what our model will try to predict. Would people show stronger emotions towards a country where a conflict arose? Or would they seem to ignore this fact and maintain their normal behavior?

Note: Remove stuff from here and instead introduce a bit of the methodology, i.e. first select tweets around a conflict's date, then identify the tweets that refer to a certain country, then get their sentiment analysis, then measure the general sentiment analysis for tweets about that country on that date, then compare for dates before and after the conflict to measure the sentimental impact.

2 Related Work

Don't know what goes here. Maybe remove this section? I also tried to find articles on this but I could not find any.

3 Data Collection

(they say it should be brief) I'm not sure about this one, since I don't know if we should skip this section or maybe here we should explain other datasets that we use. Maybe we can change the name of this section and move it 1 or 2 sections below?

4 Dataset Description

4.1 UCDP Dataset

The UCDP dataset contains data for 133012 worldwide conflicts from 1989 until 2016. Other information available for conflicts relevant for our analysis is:

- Type of violence - Categorical variable that classifies conflicts into *State-Based*, *Non-State* and *One-Sided* violence.
- Location - Coordinates, region and country are available. We are only interested in the country.
- Start and end dates - They can be used to calculate the duration of the conflict.

- Number of casualties - Both the total number and the casualties on each side involved are available. We consider only the total number.

UCDP Dataset Filtering

Since the number of worldwide active users on Twitter has increased throughout the years (Arrojo, 2015) and we are assuming that the topics covered in the tweets changed as well, we focus on conflicts that took place during 2016 - the most recent year. It is also more likely that the public reacts to certain types of violence over others, and therefore for our analysis we only consider the one-sided violence events. This type of violence represents 13.7% of the conflicts in 2016, as well as 10.7% of casualties in that year. Table 1 contains the statistics of these conflicts.

	Casualties	**Duration**	**N. per country**
mean	6.5	3 days	22.9
std	19.9	21 days	33.8
min	0	0 days	1
25%	1	0 days	2
50%	2	0 days	6
75%	5	0 days	25.5
max	324	334 days	106
count	893		

Table 1: Statistics for one-sided conflicts in 2016

As shown in the table, the number of casualties per conflict varies from 0 to 324. We assume that the public reaction will be higher for conflicts with casualties and as a result only conflicts with casualties are considered.

Table 1 also includes the number of conflicts per each country, since we are assuming that the general public is likely to lose interest on the violent events that take place in a country where conflicts are very frequent. Therefore for this analysis we consider only the countries where there were at most three one-sided violent conflicts with casualties during 2016.

The statistics for the final filtered conflicts are shown in Table 2. The result is 30 conflicts from 18 different countries.

4.2 Other Datasets

Explain here the datasets used for NER, etc (?)

	Casualties	**Duration**	**N. per country**
mean	8	0 days	1.67
std	16.9	1 days	0.69
min	1	0 days	1
25%	1	0 days	1
50%	2	0 days	2
75%	5	0 days	2
max	86	5 days	3
count	30		

Table 2: Statistics for filtered conflicts

5 Methods

(With math and description of main algorithms) I have to explain how and why I measure sentiment average and sentimental impact in those ways., for the first one I can reference a book. (Next TODO)

5.1 Language Recognition

I'm not sure about including this one!

5.2 Named Entity Recognition

For each time-frame, we do Named Entity Recognition (NER) to figure out the country it is talking about - discard all that don't mention countries of conflict.

5.3 Sentiment Analysis

For the remaining Tweets, we apply Sentiment Analysis on the text and store that information alongside the Tweets.

5.4 Sentimental Impact Measurement

We then define the Sentimental Impact as a measure that reflects the contrast between the average daily sentiment strength before and after the conflict.

6 Results and Findings

With the current schema of the report (based on the README for milestone 2) I think we would not have anything to write here.

7 Conclusions

I don't know about this and the previous section yet. The next page contains some *sample* things that I kept for reference.

8 General Instructions

Exceptions to the two-column format include full-width figures or tables (see the guidelines in Subsection 8.1).

Type of Text	Font Size	Style
report title	15 pt	bold
author names	12 pt	bold
the word “Abstract”	12 pt	bold
section titles	12 pt	bold
document text	11 pt	
captions	11 pt	
abstract text	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 3: Font guide.

8.1 The First Page

Long titles should be typed on two lines without a blank line intervening. Approximately, put the title at 2.5 cm from the top of the page, followed by a blank line, then the author’s names(s).

8.2 Sections

Headings: Do not number subsubsections.

Citations: Citations within the text appear in parentheses as (Gusfield, 1997) or, if the author’s name appears in the text itself, as Gusfield (1997). Append lowercase letters to the year in cases of ambiguity. Treat double authors as in (Aho and Ullman, 1972), but write as in (Chandra et al., 1981) when more than two authors are involved. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972). Also refrain from using full citations as sentence constituents. We suggest that instead of

> “(Gusfield, 1997) showed that ...”

you use

> “Gusfield (1997) showed that ...”

Please do not use anonymous citations

References: Gather the full set of references together under the heading **References**. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a citation as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (American Psychological Association, 1983). Use of full names for authors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the ACM *Computing Reviews* (Association for Computing Machinery, 1983).

8.3 Footnotes

Footnotes: [1]

8.4 Graphics

Illustrations: Place figures, tables, and photographs in the report near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns.

Captions: Provide a caption for every illustration; number each one sequentially in the form: “Figure 1. Caption of the Figure.” “Table 1. Caption of the Table.” captions of the figures and tables below the body.

References

Maria Jose Arrojo. 2015. *International Journal of Social Science Studies*, volume 3. Prentice-Hall, Englewood Cliffs, NJ.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

[1] This is how a footnote should appear.