Comparison of Social Media in English and Russian During Emergencies and Mass Convergence Events

Fedor Vitiugin / @vitiugin
Carlos Castillo / UPF / @chatox

upf. Universitat Pompeu Fabra Barcelona

ISCRAM 2019

Overview

Messages are collected for emergency response or research purposes in a single language.

Most previous works considered tweets in English.

Anchorage Earthquake	26,691	1,082
Ebeko Volcano Activities	2,595	258
Kerch Poly Massacre	1,267	1,358

Our hypothesis

We know there are more tweets …

but more tweets does not mean necessarily more information

We try to quantify how much is gained by doing a multi-language data collection.

Objectives

- Create event-driven parallel datasets of events across languages;
- Identify the most significant features for the comparison of tweets across languages;
- Compare the information and linguistic characteristics of these datasets.

Method overview

We focus on changes during crisis events (differences-in-differences), which can be very varied but almost invariably leave a large footprint in social media communities.

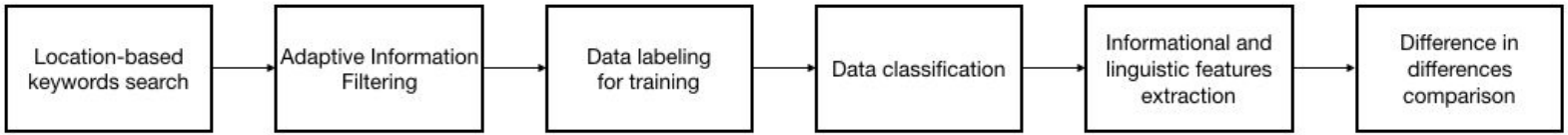
We include not only an analysis of the linguistics characteristics of messages, but also of the informativeness of messages and their sources, and virality.

Mendoza, M., Poblete, B., and Castillo, C. (2010). “Twitter Under Crisis: Can we trust what we RT?” In: Proceedings of the first workshop on social media analytics. ACM, pp. 71–79.

Tereszkiewicz, A. (2013). “Tweeting the news: a contrastive study of english and german newspaper tweets”. In: kwartalnik neofilologiczny 3.

Pipeline of research system

Keyword detection
by TF-IDF



Through TwitterAPI

	precision	recall	f1
RandomForest	**0.9022**	0.7339	**0.8079**
LinearSVM	0.8747	**0.7407**	0.8019
SGD	0.8819	0.6679	0.7550
ExtraTrees	0.8393	0.6372	0.7243

Collected data

Event	Dates	Number of tweets before filtering		Number of tweets after filtering	
		English	Russian	English	Russian
Natural disasters					
Anchorage Earthquake	01.12.18 — 03.12.18	36,865	1,263	26,691	1,082
Ebeko Volcano Activities	02.11.18 — 06.11.18	67,000	1,500	2,595	258
Man-made disasters					
Kerch Poly Massacre	18.10.18 — 20.10.18	1,850	3,350	1,267	1,358
Paris Fuel Riot	24.11.18 — 26.11.18	163,345	2,344	64,385	676
Sports events					
F1 Race in Sochi	30.09.18 — 04.10.18	333	1,650	102	189
UFC229 Khabib vs Connor	05.10.18 — 07.10.18	650	600	267	190

Results: entities

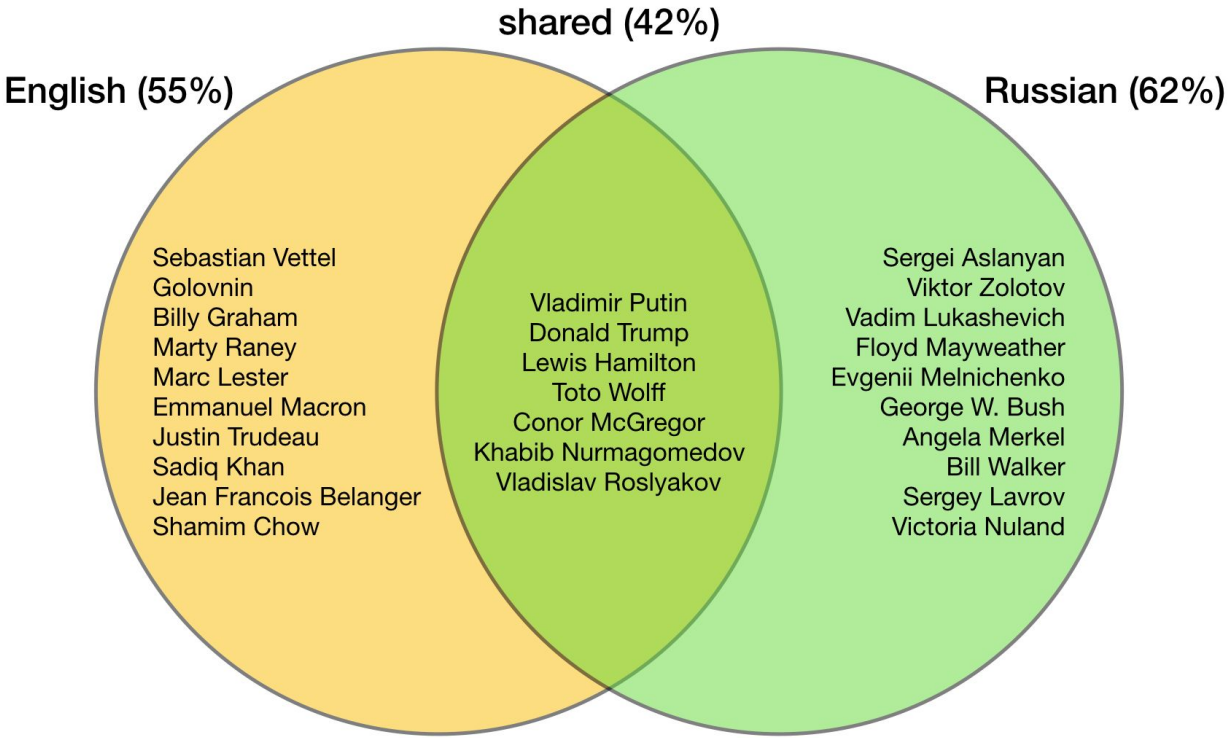
	DD_Average	DD_Median
Av. persons in English tweet	0.0941	-0.0473
Av. persons in Russian tweet	**0.1133**	**0.0599**
Av. locations in English tweet	0.6708	0.6859
Av. locations in Russian tweet	**0.9021**	**0.8418**
Av. organizations in English tweet	**0.2549***	**0.1649***
Av. organizations in Russian tweet	0.0204	0.0148
Av. unique persons in English tweet	-0.0559	-0.0999
Av. unique persons in Russian tweet	**-0.0394**	**-0.0714**
Av. unique locations in English tweet	-0.0148	**-0.0122**
Av. unique locations in Russian tweet	**-0.0131**	-0.0332
Av. unique organizations in English tweet	-0.0868	-0.0659
Av. unique organizations in Russian tweet	**-0.0118**	**-0.0286**

The '*' marks statistics for the Russian messages that differ by more than one standard deviation from the English messages.

Results: entities

	Exclusive English	English and Russian	Exclusive Russian
Persons	55 %	42 %	62 %
Locations	43 %	50 %	50 %
Organizations	86 %	9 %	76 %

Results: entities



More results

- Russian-speaking users prefer to share their own impressions, while the number of links in English tweets usually increases.

- English-speaking users are more familiar with platform mechanics than Russian-speaking users

- Russian-speaking users use Twitter as a real-time platform, to speak about what is happening now.

Results: links and citations

	DD_Average	DD_Median
Use of links (number) in English	**0.4569**	**0.4747**
Use of links (number) in Russian	0.2194	0.2649
Using of citations in English	-0.0174	-0.0312
Using of citations in Russian	**-0.0053**	**-0.0053**

Results: part of speech

Nouns and verbs as informative parts of speech for our purposes (Langacker 1987).

	DD_Average	DD_Median
POS-parsing (verbs) in English	-0.2891	-0.3262
POS-parsing (verbs) in Russian	**-0.0790***	**-0.2899**
POS-parsing (nouns) in English	-0.2247	0.0433
POS-parsing (nouns) in Russian	**0.8731***	**0.8810**

The '*' means a different of more than one standard deviation.

Results: platform mechanisms

	DD_Average	DD_Median
Type of account (name, verification) in English	**0.0328**	0,0157
Type of account (name, verification) in Russian	0.0175	**0.0164**
Dialogue (using of mentions) in English	-0.4657	**-0.4011**
Dialogue (using of mentions) in Russian	**-0.3458**	-0.4484
Spreading information (using of RT) in English	-0.3701	-0.5167
Spreading information (using of RT) in Russian	**-0.1377**	**-0.4068**
Want to trend (using of hashtags) in English	**0.2430***	**0.1821**
Want to trend (using of hashtags) in Russian	-0.0017	-0.0146

The '*' means a different of more than one standard deviation.

Results: times and numbers

	DD_Average	DD_Median
Av. time reference in English tweet	**0.0160***	-0.0417
Av. time reference in Russian tweet	-0.0100	**-0.0048**
Av. unique time reference in English tweet	-0.0178	-0.0400
Av. unique time reference in Russian tweet	**-0.0053**	**-0.0065**
Using on numbers (check of using numbers) in English	**0.3284**	**0.0855**
Using on numbers (check of using numbers) in Russian	0.0660	0.0372

The '*' means a different of more than one standard deviation.

Conclusion

The analysis of only English (or only Russian) tweets would miss a substantial amount of valuable data that can describe the effects of a crisis — detailed names of locations or new names of relevant persons.

Our analysis of named entities allows to capture:

- larger number of **locations** (in Russian) and **organizations** (in English),
- more **people** associated with an event (almost 50% of popular people are exclusive to each language).

The analysis of messages in Russian indicates an increase in the information content through a **decrease** in the **use of links** and **quotations**, a simultaneous **decrease** in the number of **verbs** and an **increase** in the number of **nouns**.

An analysis of messages in English language revealed the **activation of verified accounts**, as well as the use of **numbers** and **time references**.

Future work: classification

	CNN + Emb	LSTM + Emb	SVM	RandomForest	SGD	ExtraTrees
English	**98.73%**	96.32%	90.67%	91.35%	89.40%	91.18%
Russian	**98.04%**	94.08%	89.43%	84.70%	89.05%	90.95%
Spanish	**98.42%**	96.81%	92.70%	92.30%	92.42%	91.54%
Deutsch	**96.95%**	88.44%	89.58%	89.79%	88.85%	88.31%
average	**98.03%**	93.91%	90.60%	89.53%	89.93%	90.50%
median	**98.42%**	96.32%	90.13%	90.57%	89.23%	91.07%

Future work: data

Event	Dates	Number of tweets before filtering				Number of tweets after filtering			
		English	Russian	Spanish	German	English	Russian	Spanish	German
Anchorage Earthquake	27.11.18 — 3.12.18	16,716	1,040	2,055	155	8,834	704	1,253	109
Christchurch Massacre	15.03.19 — 18.03.19	504,441	4,877	12,553	20,702	216,955	1,631	4,412	5,228
Ethiopia Aircrush	11.03.19 — 13.03.19	41,774	1,902	7,202	735	26,597	1,314	6,346	567
Paris fuel riots	24.11.18 — 17.12.18	120,018	1,405	25,088	6,172	13,880	456	6,069	3,791

Future work

- Notability of persons
- Location types
- Organization types
- Classifiers for subjectivity
- Opinion vs fact

Questions?

fedor.vitiugin@gmail.com