Exploiting Twitter for Border Security-Related Intelligence Gathering

Jakub Piskorski\textsuperscript{1} Hristo Tanev\textsuperscript{2} Alexandra Balahur\textsuperscript{2}

Frontex\textsuperscript{1}
Research & Development
Warsaw, Poland

Joint Research Centre of the European Commission\textsuperscript{2}
Institute for the Protection and Security of the Citizen
Ispra, Italy
Outlook

- Motivation and Objective
- Retrieving tweets on migration issues posted in third countries
- Enhancing information extracted from online news on border security-related events in third countries with information extracted from Twitter
- Conclusions and Future work
Motivation

- Ever-growing amount of information is being transferred through Twitter
- Twitter might provide more up-to-date information than conventional open sources (e.g., Japan/Haiti Earthquakes, Arab Spring, Oklahoma fires)
- Information available in Twitter might not be available elsewhere
- Twitter might be used as a "sensor" for predicting threats and trends
- Exploiting Twitter might be "cheaper"
Twitter: General challenges

- **informal language used**: misspellings, non-standard abbreviations, not grammatically correct sentences, lack of punctuation

- **tweets are short**: often they do not contain well-formed ideas

- **high volume** of tweets produced on a daily basis (400 millions)

- **credibility**

- deployment of supervised ML techniques difficult due the **dynamic nature** of Twitter (Lin and Ryaboy, 2013)

- **public access is limited**
<table>
<thead>
<tr>
<th>U r flexing oh&quot;@dritchie-chad: @Nyta-Carter yea i knw.. I hv a uk visa and i also hv a shenghen..&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Okay&quot;@dritchie-chad: @Nyta-Carter yea i knw.. I hv a uk visa and i also hv a shenghen..&quot;</td>
</tr>
<tr>
<td>@iamkeezzy to gimme passport bcame wahala, he must submit ds document, change dat one reprint ds 1. Only for dem to gve hm d visa finally</td>
</tr>
<tr>
<td>@iamkeezzy nope... Actually I processed the visa 4 my brother. They kept tossing me back and forth for 3days</td>
</tr>
<tr>
<td>RT @nefertitiram: Unlike UK, Italy and France that will lazily glance thru ur docs and overlook some docs then deny you...there goes ur visa fees ?</td>
</tr>
<tr>
<td>Unlike UK, Italy and France that will lazily glance thru ur docs and overlook some docs then deny you...there goes ur visa fees ?</td>
</tr>
<tr>
<td>RT @nefertitiram: #myopinion contrary to popular belief German visa is the easiest to get! As long as you can endure all the back and forth and long delays</td>
</tr>
<tr>
<td>Their staff and security guards are so pleasant! Instead of denying u a visa, they will ask u 2 go and get any outstanding documents #german</td>
</tr>
<tr>
<td>#myopinion contrary to popular belief German visa is the easiest to get! As long as you can endure all the back and forth and long delays</td>
</tr>
<tr>
<td>@LaurasTravel hello pls cud u help me get A visa I will pay d charges</td>
</tr>
</tbody>
</table>
Objectives

Preliminary explore the usability of Twitter to gather border security-related intelligence:

- **Retrieving tweets posted in third countries with opinions and information on migration to Europe** or related issues, including exploration of sentiment analysis for improving retrieval

- **Enhancing information on crisis-related events in third countries** automatically extracted from online news with information extracted from Twitter
Retrieving tweets with opinions and information related to migration to Europe
Query generation

- QUERY: boolean expression over domain-specific keywords +
  location filter

  "passport AND visa — geocode: 9,9,700km"

- domain-specific queries obtained semi-automatically (Tanev et al., 2009)
  - specify seed terms, e.g. EU countries, target for irregular migration
  - automatically extract co-occurrence patterns, i.e., word uni-
    and bigrams, from a 1.4 million tweet corpus from Africa and Middle East, e.g. "moving to"
  - manually refine co-occurrence patterns and use them as keywords in the query
### Query generation

<table>
<thead>
<tr>
<th>Query</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(“prison in” OR “staying back” OR passport OR “illegal in” OR “living in” OR “working in” OR “lives in” OR “lived in” OR “moving to” OR visa OR “house in” OR “as a country”) AND (Italy OR UK OR Germany OR Austria OR Spain OR Belgium OR Holland OR France OR Portugal OR Turkey OR Greece)</td>
<td>AND (Italy OR UK OR Germany OR Austria OR Spain OR Belgium OR Holland OR France OR Portugal OR Turkey OR Greece)</td>
</tr>
<tr>
<td>(immigrate OR emigrate OR immigrant OR emmigrant OR immigrated OR immigrants OR emmigrants OR emmigrated) AND (Italy OR UK OR Germany OR Austria OR Spain OR Belgium OR Holland OR France OR Portugal OR Turkey OR Greece)</td>
<td>AND (Italy OR UK OR Germany OR Austria OR Spain OR Belgium OR Holland OR France OR Portugal OR Turkey OR Greece)</td>
</tr>
<tr>
<td>visa AND passport</td>
<td>visa AND passport</td>
</tr>
<tr>
<td>(fake OR valid) AND (passport OR visa)</td>
<td>(fake OR valid) AND (passport OR visa)</td>
</tr>
</tbody>
</table>
Sentiment Analysis: Classification

- classification model trained using Support Vector Machines Sequential Minial Optimization (SVM SMO) with linear kernel on 90K of training data (from different domains)

- boolean features: presence/absence of unigrams, bigrams and unigrams + bigrams

- tweet preprocessing to generalize n-gram features

- no specific language analysis components used

- accuracy over 80% (Balahur, 2013)
Sentiment Analysis: Tweet Pre-processing

Whoooop Whoooop!!! Austria will never be the same RT @DonBag101 Visa approved :) yay 22 days to Europe #EpicDubs_Crew #Europe2013

- **Step 1:** **Repeated punctuation sign normalization:**
  
  Whoooop Whoooop **MULTIEXCLAMATION** Austria will never be the same RT @DonBag101 Visa approved :) yay 22 days to Europe #EpicDubs_Crew #Europe2013

- **Step 2:** **Emoticon replacement:**

  Whoooop Whoooop **MULTIEXCLAMATION** Austria will never be the same RT @DonBag101 Visa approved **POSITIVE** yay 22 days to Europe #EpicDubs_Crew #Europe2013
Sentiment Analysis: Tweet Pre-processing

- **Step 3: Lower casing and Tokenization:**

  `whoooop whoooop multiexclamation austria will never be the same rt @donbag101 visa approved positive yay 22 days to europe #epicdubs_crew #europe2013`

- **Step 4: Slang replacement:**

  `whoooop whoooop multiexclamation austria will never be the same rt @donbag101 visa approved positive great 22 days to europe #epicdubs_crew #europe2013`

- **Step 5: Word normalization (Roget’s thesaurus):**

  `whoop whoop multiexclamation austria will never be the same rt @donbag101 visa approved positive great 22 days to europe #epicdubs_crew #europe2013`
Sentiment Analysis: Tweet Pre-processing

- **Step 6:** Affective word matching (General Inquirer, LIWC, MicroWNOp):
  
  h-positive h-positive multiexclamation austria will never be the same rt @donbag101 visa positive positive h-positive 22 days to europe #epicdubs_crew #europe2013

- **Step 7:** Modifier word matching:
  
  h-positive h-positive multiexclamation austria will negator be the same rt @donbag101 visa positive positive h-positive 22 days to europe #epicdubs_crew #europe2013

- **Step 8:** User and topic labeling:
  
  h-positive h-positive multiexclamation austria will negator be the same rt user visa positive positive h-positive 22 days to europe topic topic
Whooop Whooop!!! Austria will never be the same RT @DonBag101 Visa approved :) yay 22 days to Europe #EpicDubs_Crew #Europe2013

h-positive h-positive multiexclamation austria will negator be the same rt user visa positive positive h-positive 22 days to europe topic topic
Experiments

- Submitted queries with location restriction to Nigeria to Twitter Search API

- After deleting duplicates 156 tweets were obtained and tagged as: strictly relevant, loosely relevant and irrelevant

- Evaluation:
  - all tweets: 42% strictly relevant, 72% strictly or loosely relevant, 28% irrelevant
  - only positive and negative tweets (77): 47% strictly relevant
<table>
<thead>
<tr>
<th>Experiments: example of retrieved tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Get ur VISA ASAP Traveling Agent: we are here 4 U.UK visa, USA Visa, Shegen Visa and Canada Visa etc. for more info 0706630...</td>
</tr>
<tr>
<td>Embassy begins head count of Nigerians living in Germany <a href="http://j.mp/mAXcNI">http://j.mp/mAXcNI</a> ...</td>
</tr>
<tr>
<td>I wish I lived in UK, their lifestyle is similar to mine</td>
</tr>
<tr>
<td>If they gimme a UK visa right now, I’ll prolly burn it. It’s just like me saying I want to go to Obalende to chill. FTBS.</td>
</tr>
<tr>
<td>Immigrants in Greece suffer hell in the hands of Greek people and it’s government...</td>
</tr>
<tr>
<td>What has Britain done in recent times to aid Nigeria? Zero! D revenue generated for the UK consular in Nigeria 4 visa is still...</td>
</tr>
<tr>
<td>2000 dollars for a date uno. I could a fake passport and be headed to Uk/Canada or America with that</td>
</tr>
<tr>
<td>I’m not in France yet i’m still in Africa but when I have enough money for myself I will surely be moving to France</td>
</tr>
<tr>
<td>I thought Uk visa application waiting time never exceeds 3 weeks. This is so frustrating!!!</td>
</tr>
<tr>
<td>If I still lived in Holland I’d probably be smoking weed everyday with randa and hannni</td>
</tr>
</tbody>
</table>
Experiments

- Submitted queries with location restriction to some other Central and Western African countries (including: Burundi, Central African Republic, Congo, Democratic Republic of Congo, Gambia, Ghana, Guinea Bissau, Ivory Coast, Liberia, Mali, Mauritania, Niger, Rwanda, Sierra Leone, Senegal) resulted in retrieval of **circa 150 tweets only**.

→ French is the most prevalent language in these countries
Event-related Tweet Retrieval

- submit several boolean queries which are primarily based on keywords and key phrases extracted from the first sentences of the news articles reporting on the event (Tanev et al., 2012)
- build term vector for the event news story and the tweets
- compute similarity: projection of the tweet vector on the story vector
- filter out tweets below empirically set threshold
### Example: Event-related tweets

<table>
<thead>
<tr>
<th>Tweet</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reports Turkish Police Using Rubber Bullets On Protesters: via @youtube <a href="http://www.youtube.com/watch?v=js6FVAwItDU">http://www.youtube.com/watch?v=js6FVAwItDU</a></td>
<td></td>
</tr>
<tr>
<td>RT @b judah: As Erdogan tear gasses Taksim again - reposting my piece - Why Erdogan Doesnt Get It. He’s fighting the Deep State.</td>
<td>Are illegal surveillance, water cannons, rubber bullets and tear gas the new hallmarks of a free and democratic society?? How free are you?</td>
</tr>
<tr>
<td>RT @RT.com: Reports of dozens of injured in clashes as Turkish police blocked streets leading to Taksim</td>
<td>RT @RT.com: DETAILS: Tear gas, water cannon in Gezi Park and Taksim <a href="http://rt.com/news/gezi-ultimatum-erdogans-tear-763/">http://rt.com/news/gezi-ultimatum-erdogans-tear-763</a></td>
</tr>
</tbody>
</table>
Twitter Link Ranking

\[
\text{score}(\text{URL}) = ((\text{Mentions} - 1) + \text{Retweets} \times 1, 3 + \text{Favorited} \times 4) \\
\times (\text{InConv} + 2 \times \text{ReplyTo} + 1)
\]

- **Mentions** - number of mentions of the URL
- **Retweet** - number of tweets with the URL
- **InConv** - number of mentions in conversations
- **Favorited** - number of times the URL was favorited
- **ReplyTo** - number of tweets which replied to tweets, mentioning the URL

J. Piskorski, H. Tanev and A. Balahur
Exploiting Twitter for Intelligence Gathering
Example: Event enhanced with Twitter Links

MORE than 4,000 people have fled at night from a border town into Uganda after rival rebel factions fought in the Democratic Republic of Congo. The violence followed the sacking of the political leader of the M23 rebel group, Jean-Marie Runiga, on Thursday. Mr. ...

**Slot Name** | **Value**
---|---
TYPE | Crisis
SUBTYPE | Humanitarian Crisis
DATE | 2013-03-01
LOCATION | Congo
COUNTRY | Democratic Republic of the Congo
COD | displaced | people (4000)
cause | Armed Conflict
SNIPPET | MORE than 4,000 people have fled at night from a border town into Uganda ...

twitterlinks: [https://abcnews.go.com](https://abcnews.go.com), [www.bbc.co.uk](http://www.bbc.co.uk), [www.nytimes.com](http://www.nytimes.com)
Experiments & Evaluation

- Two test corpora were created consisting of randomly selected events extracted automatically from the news:
  - **35** security-related events in Africa and Middle East in the period of 15 April - 7 May 2013
  - **28** security-related events in the countries in the proximity of Eastern EU external borders (e.g., Russia, Ukraine, Turkey, Armenia, Georgia, etc.) in the period of 20 April - 20 May 2013

- Mainly focused of crisis events relevant for third country situation monitoring: unrests, armed conflicts, displacement of people, violent incidents, man-made disasters.

- On average 4-5 Twitter links could be extracted for each event in the test corpus
## Experiments & Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Africa and Middle East</th>
<th>Eastern Borders Proximity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Events for which Twitter links could be extracted</td>
<td>88.6%</td>
<td>75%</td>
</tr>
<tr>
<td>Links that point to <strong>relevant information</strong></td>
<td>95.3%</td>
<td>87.6%</td>
</tr>
<tr>
<td>Links that point to <strong>news sites</strong></td>
<td>86.7%</td>
<td>82.5%</td>
</tr>
<tr>
<td>Links that point to <strong>non-news sites</strong></td>
<td>13.3%</td>
<td>17.5%</td>
</tr>
</tbody>
</table>
Conclusions

- Exploration of methods to exploit Tweeter to:
  - retrieve tweets posted in third countries with opinions and information related to migration to Europe
  - enhance news information on events in third countries with "Twitter links"

- Methods are highly language-independent

- Preliminary evaluation results are promising
Future work

- Further refining tweet retrieval and twitter link extraction processes
- Evaluation on a larger chunk of data
- Tackling multilinguality
- Comparison of the intensification of Twitter messages on migration to Europe versus "real data" on irregular migration to Europe
- Evaluation of usability/complementarity in an operational environment over longer period of time
References for further reading


A. Balahur. *Sentiment Analysis in Social Media* In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 120128, Atlanta, Georgia, 2013.


Thank You!