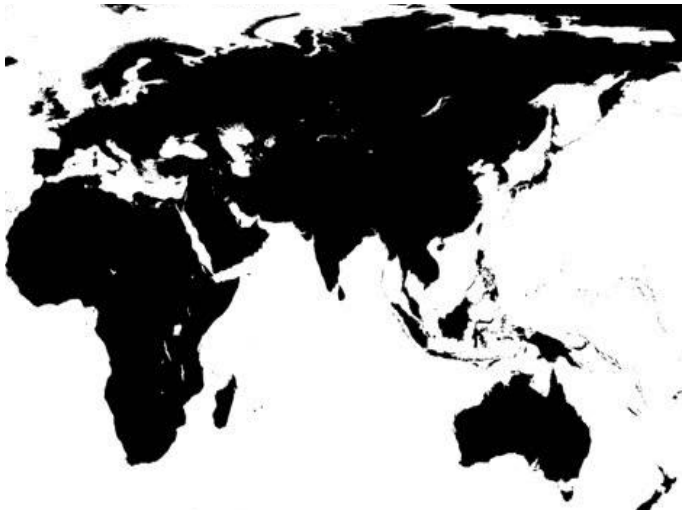
ANALYZING TEXT OF CONFLICT REPORTS

RECAP: THE DATA

ACLED
Bringing clarity to crisis

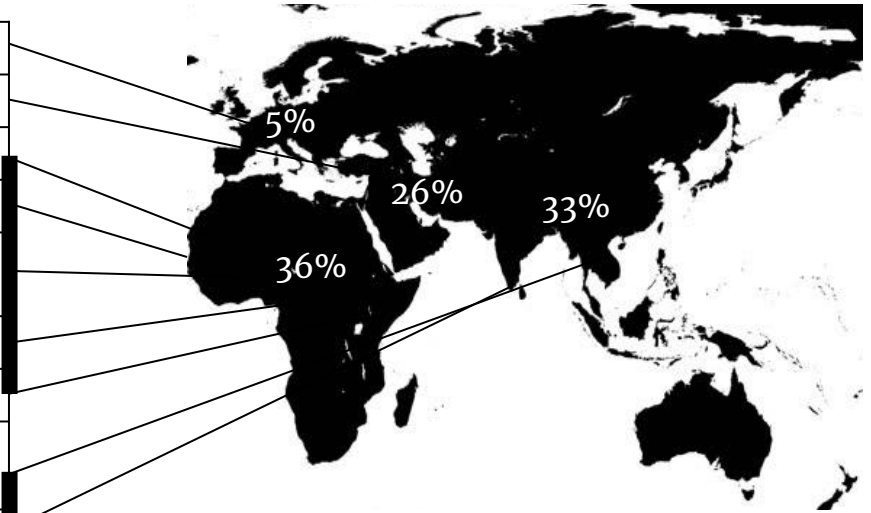
disaggregated conflict analysis and crisis mapping project.



Source @ acleddata.com

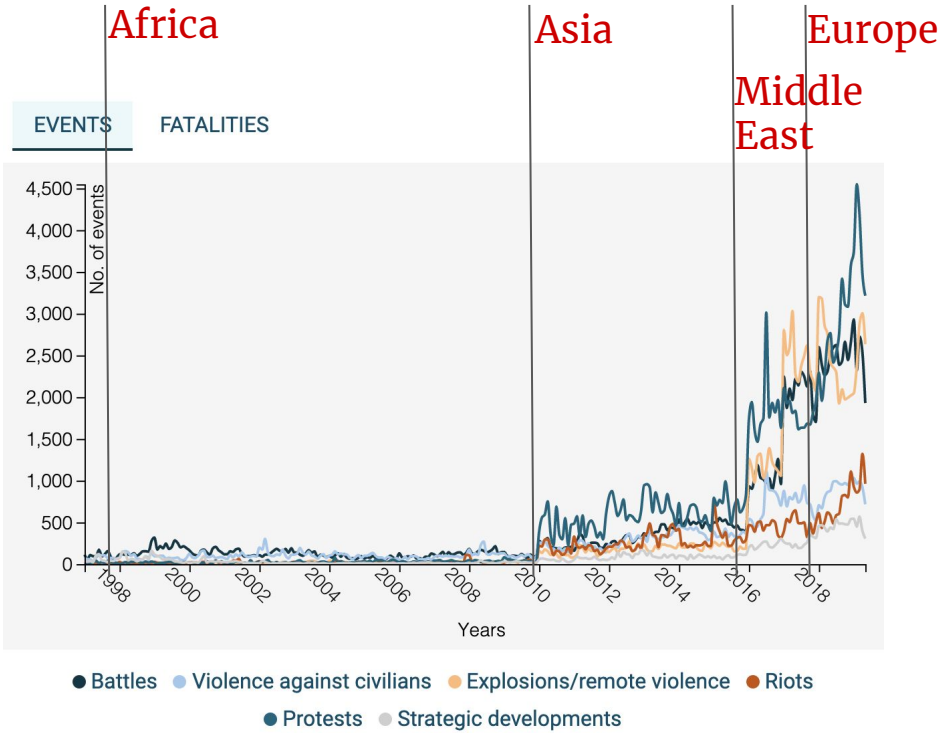
RECAP: THE DATA

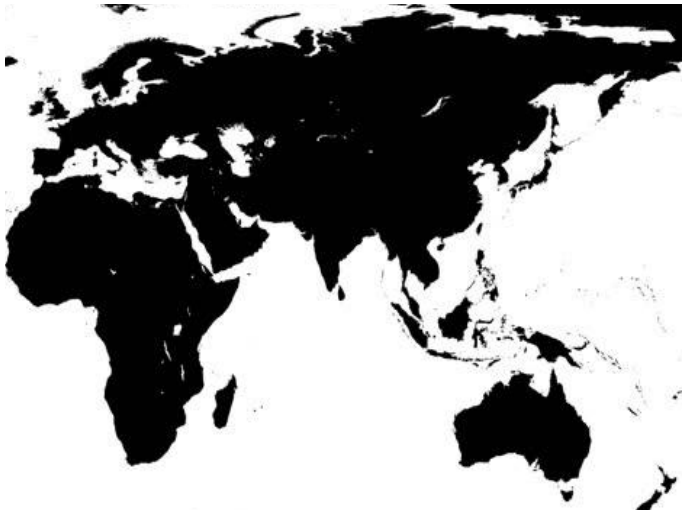
Europe	26315	5.17
Middle East	134805	26.48
Eastern Africa	62995	12.37
Northern Africa	42195	8.29
Western Africa	31515	6.19
Middle Africa	25195	4.95
Southern Africa	19085	3.75
South-Eastern Asia	25502	5.01
Southern Asia	141550	27.8



Source @ acleddata.com

RECAP: THE DATA

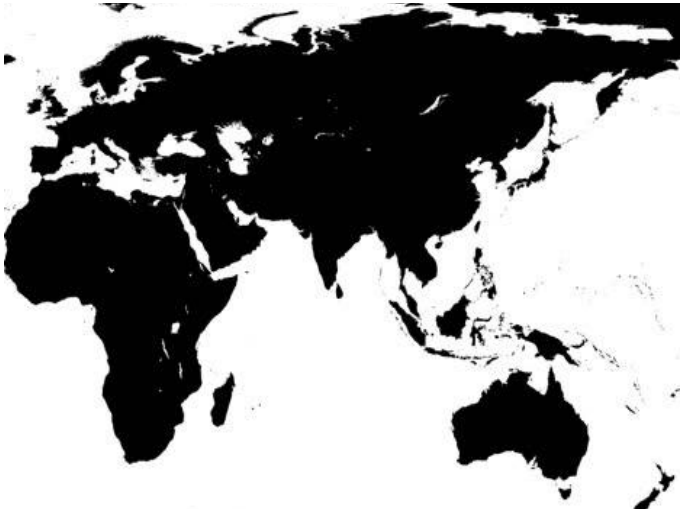




Source @ acleddata.com

RECAP: TYPES OF EVENTS

Abduction/forced disappearance	Agreement
5058	1503
Air/drone strike	Armed clash
41131	110362
Arrests	Attack
3226	61388
Change to group/activity	Chemical weapon
4740	110
Disrupted weapons use	Excessive force against protesters
3830	2490
Government regains territory	Grenade
7111	3481
Headquarters or base established	Looting/property destruction
959	5466
Mob violence	Non-state actor overtakes territory
19444	4228
Non-violent transfer of territory	Other
3169	2184
Peaceful protest	Protest with intervention
126411	10239
Remote explosive/landmine/IED	Sexual violence
23238	1767
Shelling/artillery/missile attack	Suicide bomb
41151	1638
Violent demonstration	
24833	

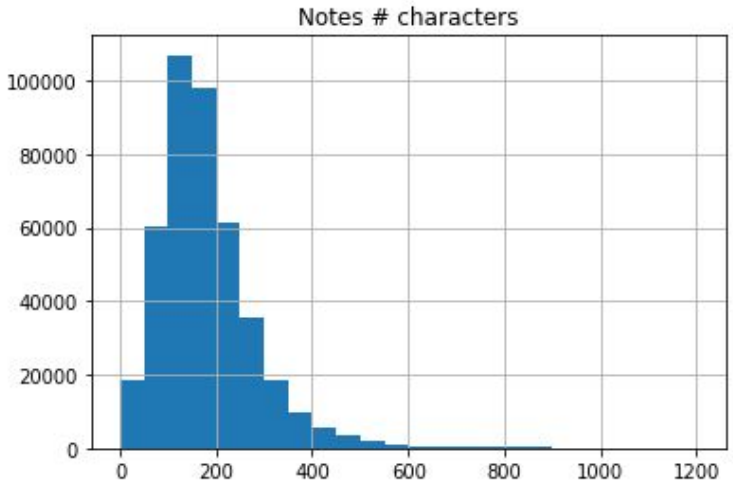


Source @ acleddata.com

RECAP: THE TEXT

509,157 Observations

Common pre-processing: Keeping those observations with > 100 characters



NOTES examples:

“26th Feb 2001- BBC Mon-Large military offensive all over the country sees 9 soldiers and 6 GIA killed” (101 characters)

“A 40-year-old repentant who answered to the name of Hamid Doghman was assassinated this past Tuesday 4 March at about 2000 hours not far from downtown Zemmouri 16 kilometres east of Boumerdes by a militant group made up of between four and six elements. The attack targeted this ex-Islamist who had signed his repentance in August 1995.”

Source @ acleddata.com

GOAL 1:
Identify contesting parties

GOAL 2:
Relationship extraction

GOAL 3:
Topic analysis extraction

GOAL 1:
Identify contesting parties

GOAL 2:
Relationship extraction

GOAL 3:
Topic analysis extraction

ACTOR1	INTER1	ACTOR2	INTER2	INTERACTION	NOTES
GIA: Armed Islamic Group	Rebel Groups (2)	Military Forces of Algeria (1999-)	State Forces (1)	12	26th Feb 2001- BBC Mon-Large military offensive all over the country sees 9 soldiers and 6 GIA killed
Unidentified Armed Group (Algeria)	Political Militias (3)	Civilians (Algeria)	Civilians (7)	37	A 40-year-old repentant who answered to the name of Hamid Doghman was assassinated this past Tuesday 4 March at ...

GOAL 1:
Identify contesting parties

Initially:

- Predict specific actors (e.g. Syria's Military, Boko Haram, village)
- Use SpaCy or Stanford NER to perform regular Name Entity Recognition and POS inputs.
- Predict "actor1" and "actor2".
- Use neural networks

Challenges:

- Difficulty tagging organizations accurately.
- Actor labels often incorporated more information than notes.
- Complexity of predicting two variables (actors) at the same time

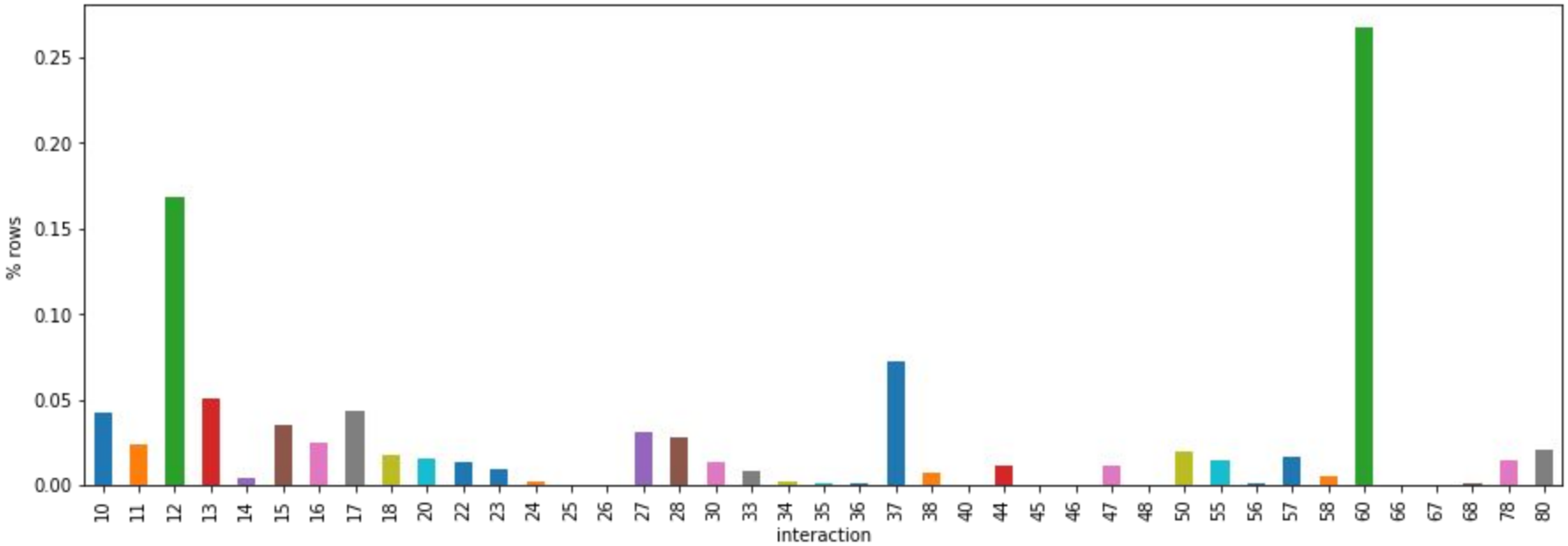
ACTOR1	INTER1	ACTOR2	INTER2	INTERACTION	NOTES
GIA: Armed Islamic Group	Rebel Groups (2)	Military Forces of Algeria (1999-)	State Forces (1)	12	26th Feb 2001- BBC Mon-Large military offensive all over the country sees 9 soldiers and 6 GIA killed
Unidentified Armed Group (Algeria)	Political Militias (3)	Civilians (Algeria)	Civilians (7)	37	A 40-year-old repentant who answered to the name of Hamid Doghman was assassinated this past Tuesday 4 March at ...

GOAL 1:
Identify contesting parties

New approach:

- Predict “Interaction” (41 categories) using a Neural Network with 2 hidden layers (size 128 and 100).
- Use pre-trained SpaCy word2vec embeddings as input. Provides vector of 300 dimensions for complete sentences.

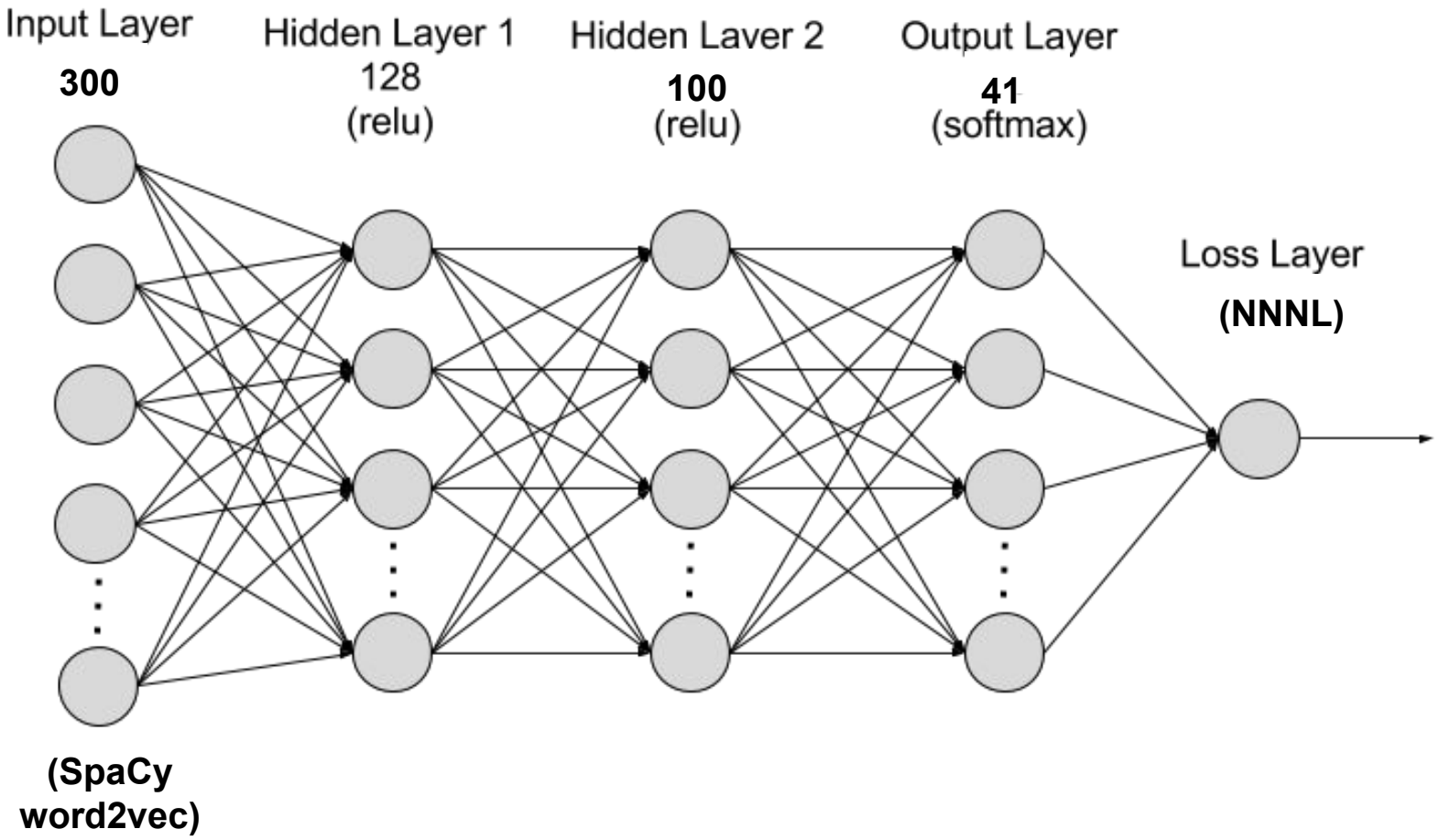
10- SOLE MILITARY ACTION	26- REBELS VERSUS PROTESTERS	47- COMMUNAL MILITIA VERSUS CIVILIANS
11- MILITARY VERSUS MILITARY	27- REBELS VERSUS CIVILIANS	48- COMMUNAL MILITIA VERSUS OTHER
12- MILITARY VERSUS REBELS	28- REBELS VERSUS OTHERS	50- SOLE RIOTER ACTION
13- MILITARY VERSUS POLITICAL MILITIA	30- SOLE POLITICAL MILITIA ACTION	55- RIOTERS VERSUS RIOTERS
14- MILITARY VERSUS COMMUNAL MILITIA	33- POLITICAL MILITIA VERSUS POLITICAL MILITIA	56- RIOTERS VERSUS PROTESTERS
15- MILITARY VERSUS RIOTERS	34- POLITICAL MILITIA VERSUS COMMUNAL MILITIA	57- RIOTERS VERSUS CIVILIANS
16- MILITARY VERSUS PROTESTERS	35- POLITICAL MILITIA VERSUS RIOTERS	58- RIOTERS VERSUS OTHERS
17- MILITARY VERSUS CIVILIANS	36- POLITICAL MILITIA VERSUS PROTESTERS	60- SOLE PROTESTER ACTION
18- MILITARY VERSUS OTHER	37- POLITICAL MILITIA VERSUS CIVILIANS	66- PROTESTERS VERSUS PROTESTERS
20- SOLE REBEL ACTION	38- POLITICAL MILITIA VERSUS OTHERS	67- PROTESTERS VERSUS CIVILIANS
22- REBELS VERSUS REBELS	40- SOLE COMMUNAL MILITIA ACTION	68- PROTESTERS VERSUS OTHER
23- REBELS VERSUS POLITICAL MILIITA	44- COMMUNAL MILITIA VERSUS COMMUNAL MILITIA	78- OTHER ACTOR VERSUS CIVILIANS
24- REBELS VERSUS COMMUNAL MILITIA	45- COMMUNAL MILITIA VERSUS RIOTERS	80- SOLE OTHER ACTION
25- REBELS VERSUS RIOTERS	46- COMMUNAL MILITIA VERSUS PROTESTERS	


0.25
0.20
0.15
0.10
0.05
0.00
% rows
10
11
12
13
14
15
16
17
18
20
22
23
24
25
26
27
28
30
33
34
35
36
37
38
40
44
45
46
47
48
50
55
56
57
58
60
66
67
68
78
80
interaction

GOAL 1:
Identify contesting parties

Pipeline:

- Pre-processing:
  - drop notes with <100 characters and three strange codes
  - Split data: 75% oldest news training (~310K), random 12.5% split test and validation (~52K each)
- Train Neural network:
  - Apply SpaCy word2vec to each notes
  - 2 epochs through entire training data
  - Learning rate = 0.001; Adam optimizer
  - Each 80,000 observations check in validation and save best model
- Use a simple logistic regression “neural network” as baseline (HW1)

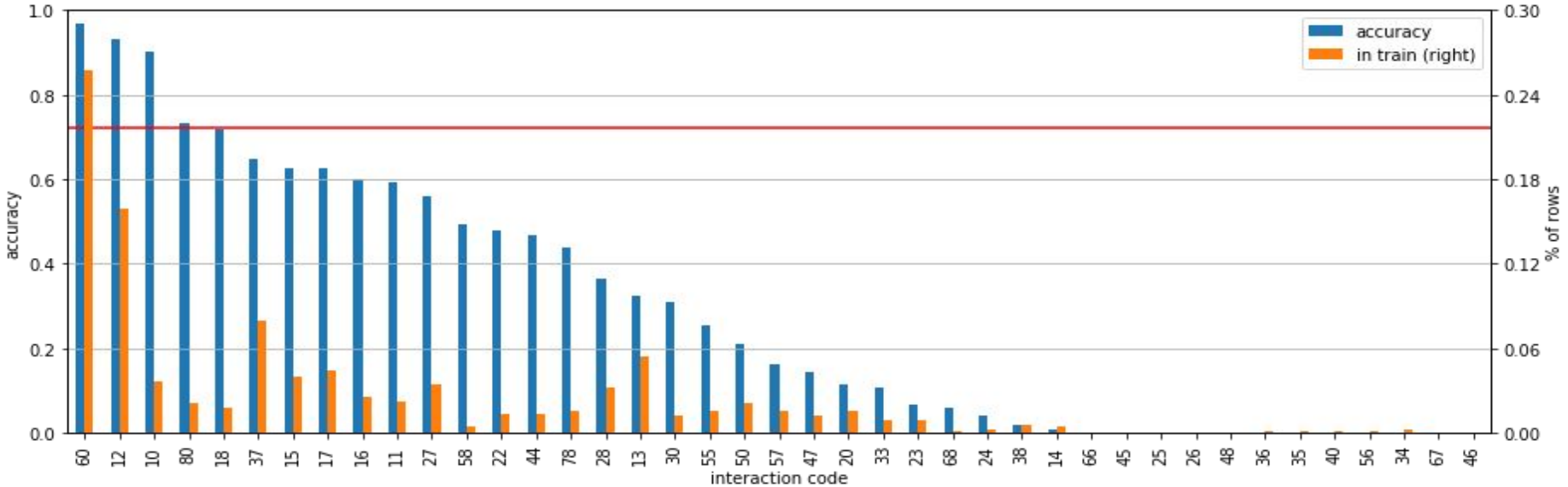

Input Layer
300
Hidden Layer 1
128
(relu)
Hidden Laver 2
100
(relu)
Output Layer
41
(softmax)
Loss Layer
(NNNL)
(SpaCy
word2vec)

Results:

- Test error: 0.2774
- BL model: 0.4911

- Top 5:
  - 60- SOLE PROTESTER ACTION
  - 12- MILITARY VERSUS REBELS
  - 10- SOLE MILITARY ACTION
  - 80- SOLE OTHER ACTION
  - 18- MILITARY VERSUS OTHER

- While number of training observations was important for accuracy, there is more happening.
- Words or actions specific to categories: “Coalitions”, “Protestors”, “Established base”



- 60- SOLE PROTESTER ACTION

“A protest was staged in Colombo on 26 October 2017 demanding the release of IUSF activists and other students in remand custody.”

- 12- MILITARY VERSUS REBELS

“MoD reports Afghan army conducted military operations across 17 provinces; killing 36 suspected Taliban militants. Fatalities split across 15 provinces, with some specific events listed in article. 21 fatalities coded in this series while 15 previously coded in other events. 15 Events in total as operations, with Nangarhar and Kunduz were already coded separately.”

- 10- SOLE MILITARY ACTION

“Military forces established an operational base in Kazimiya, in order to better control this area on Lake Tanganyika.”

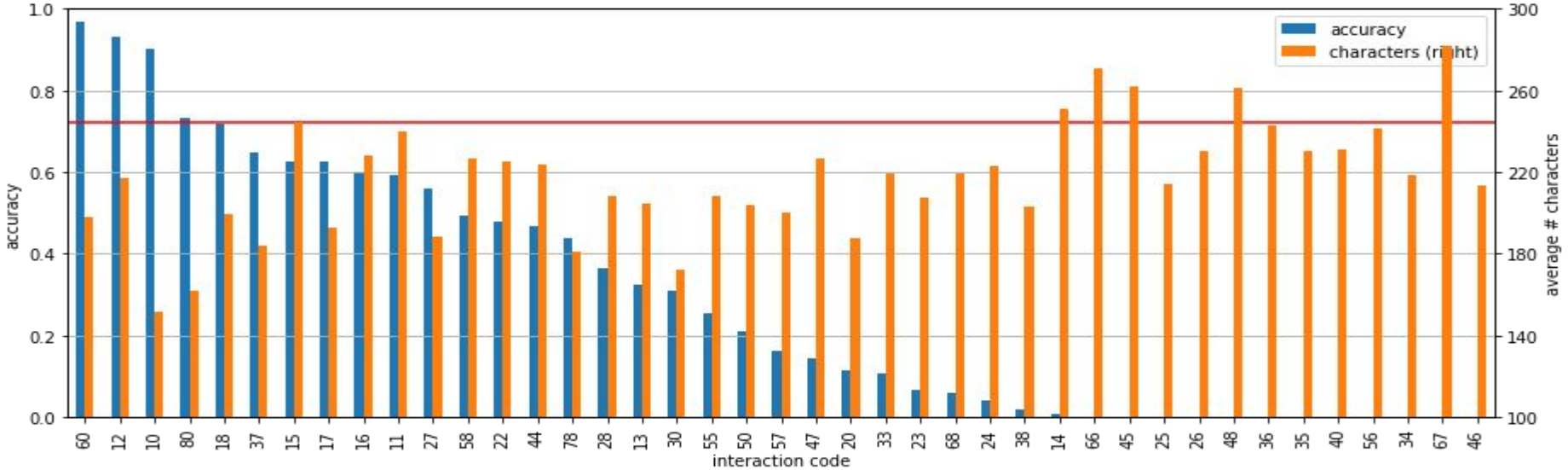
- 80- SOLE OTHER ACTION

“The Saudi-led coalition carried out three air raids on the Atias mountain and two air raids on the Nashr area in the Sirwah district, Marib governorate. No casualties were reported but private property was damaged.”

- 18- MILITARY VERSUS OTHER

“Pro-Houthi forces claim to have shelled Saudi soldiers in Raqabat-sudais and Makhroq, Najran. No injuries reported.”

- # characters doesn’t seem to help that much
- Regression of 'correct_pred ~ C(interaction_code) + C(year) + characters' shows each additional character *decreases* accuracy in 0.03 pp.
- News from 2018 also had higher accuracy.



- Notes and labels were not super clean.

GOAL 1:
Identify contesting parties

GOAL 2:
Identify relationship between parties

GOAL 3:
Topic analysis extraction

- Pre-processing
- LSTM
- Bidirectional LSTM
- Caveats
- Ongoing work

GOAL 2:
Identify relationship between parties

Initial proposal: ***Relationship extraction***
A attacks B or *B attacks A*
But all of our cases where we know directionality go in the same direction

Implementation: Multi Classification task
What is the action that A inflicts on B
Classes: 333 specific kinds of attacks

- ➔ Pre-processing
- ➔ LSTM
- ➔ Bidirectional LSTM
- ➔ Caveats
- ➔ Ongoing work

1. Subset data to cases where we knew who the aggressor is:
   a. Attacks on civilians

2. Use SpaCy POS tagger to find all verbs that appear in the corpus

3. Manually identify verbs related to violent action; label cases according to that list.
4. Update string

GOAL 2:
Identify relationship between parties

- ➔ Pre-processing
- ➔ LSTM
- ➔ Bidirectional LSTM
- ➔ Caveats
- ➔ Ongoing work

Challenge: No labelled data.

1. Subset data to cases where we knew who the aggressor is:

‘two VICTIM injured while farming in a northern village by AGGRESSOR’

‘AGGRESSOR injured after having shelled a VICTIM ’

GOAL 2:
Identify relationship between parties

- ➔ Pre-processing
- ➔ LSTM
- ➔ Bidirectional LSTM
- ➔ Caveats
- ➔ Ongoing work

Challenge: No labelled data.

4. Identify all the ways in which aggressors are mentioned
Ex: ‘AQAP: Al Qaeda in the Arabian Peninsula’
[‘AQAP’, ‘Al Qaeda’, ‘AQAP: Al Qaeda in the Arabian Peninsula’]

Ex: ‘Unidentified Armed Forces’
[‘Unidentified Armed Forces’ , ‘Unidentified Armed Group’, ‘Unidentified Forces’, (...)]
1238 categories in the end

5. Identify all the ways in which victims are mentioned
[‘civilians’, ‘citizens’, ‘families’, ‘villagers’, ...]

GOAL 2:
Identify relationship between parties

- ➔ Pre-processing
- ➔ LSTM
- ➔ Bidirectional LSTM
- ➔ Caveats
- ➔ Ongoing work

Challenge: No labelled data.

- → Pre-processing
- → LSTM
- → Bidirectional LSTM
- → Caveats
- → Ongoing work

4. Identify all ways in which aggressors are mentioned
Ex: 'AQAP: Al Qaeda in the Arabian Peninsula'
['AQAP', 'Al Qaeda', 'AQAP: Al Qaeda in the Arabian Peninsula']

Ex: 'Unidentified Armed Forces'

10	1 February: 7 citizens of a farming estate (Ha...	1 february: 7 VICTIM of a farming estate (haou...
11	16 February: killing of two families from the ...	16 february: killing of two VICTIM from the el...
12	6 April: 15 people killed by an armed group in...	6 april: 15 VICTIM killed by an AGGRESOR in an...

['civilians', 'citizens', families', 'villagers', ...]

6. Update string

1. Total Data — 509,157
2. > 100 characters — 415,462
3. We can identify directionality — 50,671
4. Have 'victim' & 'aggressor' strings — 8,294
5. Have 'victim' OR 'aggressor' strings — 34,556

Recurrent Neural Network approach

Results with:

- Optimization: Adam
  - Adjusting LR for every iteration
  - Default alpha
- Loss:
  - Negative Logistic Log Likelihood
- N_epochs:
  - 15.
- HidDim:
  - 30
- Embeddings:
  - 10

- ➔ Pre-processing
- ➔ LSTM
- ➔ Bidirectional LSTM
- ➔ Caveats
- ➔ Ongoing work

```
epoch: 0; accuracy: 0.6651672757942086; NLLLoss: 5548.86279296875
epoch: 1; accuracy: 0.6634804610626932; NLLLoss: 4939.9013671875
epoch: 2; accuracy: 0.6854090525723925; NLLLoss: 4709.9404296875
epoch: 3; accuracy: 0.6927185830756255; NLLLoss: 4528.18359375
epoch: 4; accuracy: 0.7084621872364352; NLLLoss: 4319.44677734375
epoch: 5; accuracy: 0.7180208040483553; NLLLoss: 4172.93505859375
epoch: 6; accuracy: 0.7242057913972448; NLLLoss: 4076.931884765625
epoch: 7; accuracy: 0.7242057913972448; NLLLoss: 4035.460205078125
epoch: 8; accuracy: 0.7197076187798707; NLLLoss: 4047.114501953125
epoch: 9; accuracy: 0.7374191734607816; NLLLoss: 4034.38330078125
epoch: 10; accuracy: 0.726454877705932; NLLLoss: 4092.075927734375
epoch: 11; accuracy: 0.7272982850716896; NLLLoss: 4023.69873046875
epoch: 12; accuracy: 0.7382625808265393; NLLLoss: 4057.7822265625
epoch: 13; accuracy: 0.7292662355917908; NLLLoss: 4008.925048828125
epoch: 14; accuracy: 0.733764408209165; NLLLoss: 4051.241455078125
epoch: 15; accuracy: 0.7329210008434074; NLLLoss: 4007.497314453125
```

Best model evaluated on test:

73.42%

Recurrent Neural Network approach

Results with:

- Optimization: Adam
  - Adjusting LR for every iteration
  - Default alpha
- Loss:
  - Negative Logistic Log Likelihood
- N_epochs:
  - 15.
- HidDim:
  - 30
- Embeddings:
  - 10

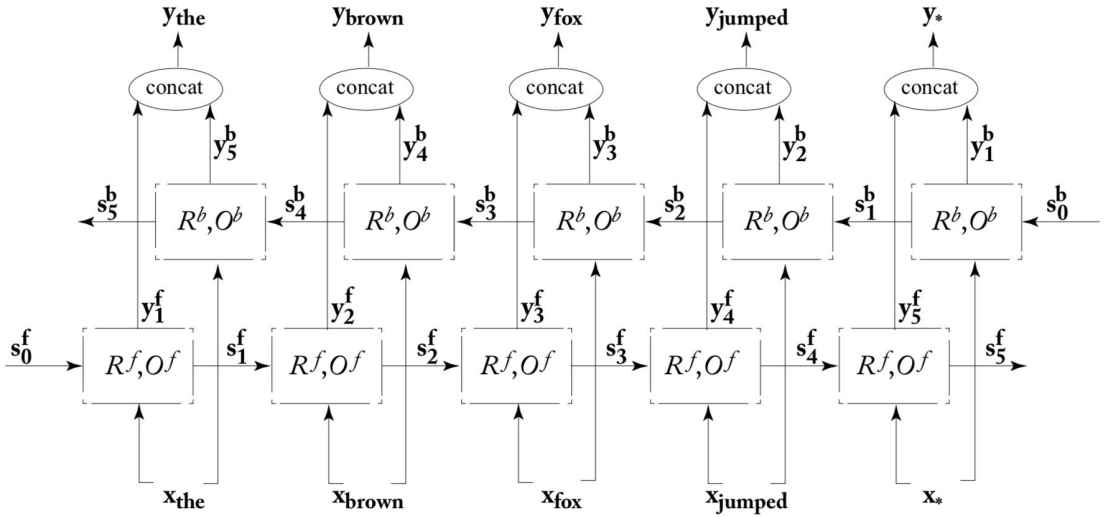
- ➔ Pre-processing
- ➔ LSTM
- ➔ Bidirectional LSTM
- ➔ Caveats
- ➔ Ongoing work

```
epoch: 0; accuracy: 0.6671296296296296; NLLLoss: 5698.32666015625
epoch: 1; accuracy: 0.7212962962962963; NLLLoss: 4979.095703125
epoch: 2; accuracy: 0.7388888888888889; NLLLoss: 4784.62890625
epoch: 3; accuracy: 0.7363425925925926; NLLLoss: 4754.27490234375
epoch: 4; accuracy: 0.7379629629629629; NLLLoss: 4764.02001953125
epoch: 5; accuracy: 0.7391203703703704; NLLLoss: 4830.21923828125
epoch: 6; accuracy: 0.7342592592592593; NLLLoss: 4882.53857421875
epoch: 7; accuracy: 0.7361111111111112; NLLLoss: 4978.2236328125
epoch: 8; accuracy: 0.7354166666666666; NLLLoss: 5126.94775390625
epoch: 9; accuracy: 0.7368055555555555; NLLLoss: 5049.662109375
epoch: 10; accuracy: 0.7261574074074074; NLLLoss: 5127.7294921875
epoch: 11; accuracy: 0.724537037037037; NLLLoss: 5265.04248046875
epoch: 12; accuracy: 0.7296296296296296; NLLLoss: 5178.81640625
epoch: 13; accuracy: 0.7259259259259259; NLLLoss: 5345.2041015625
epoch: 14; accuracy: 0.7259259259259259; NLLLoss: 5290.7275390625
```

Best model evaluated on test:

74.35%

Neural Network Methods for Natural Language Processing



GOAL 2:
Identify relationship between parties

- ➔ Pre-processing
- ➔ LSTM
- ➔ Bidirectional LSTM
- ➔ Caveats
- ➔ Ongoing work

Accuracy from test
LSTM 73.42%
Bi-dir 74.35%

- Pre-processing
- LSTM
- Bidirectional LSTM
- Caveats
- Ongoing work

- Tune both approaches changing
  - Optimizer
  - Learning Rate
  - Alpha
  - Weight decay

- See if our predictions are balanced across:
  - Years
  - Countries / regions
  - Types of conflicts

GOAL 2:
Identify relationship between parties

GOAL 1:
Identify contesting parties

GOAL 2:
Identify contesting parties

GOAL 3:
Topic analysis extraction

GOAL 1:
Identify contesting parties

GOAL 2:
Identify relationship between parties

GOAL 3:
Topic analysis extraction

Motivation

1. Extract underlying themes & topics in the ACLED data

2. Use these to identify similarities in conflicts around the world

3. Understand distributions by geography & trends

GOAL 3:
Topic analysis extraction

- ➔ Goals
- ➔ Recent Literature
- ➔ Pre-processing
- ➔ LDA
- ➔ Visualizations
- ➔ Ongoing/future work

- ➔ Goals
- ➔ Recent Literature
- ➔ Pre-processing
- ➔ LDA
- ➔ Visualizations
- ➔ Ongoing/future work

Tarik Altuncu, M et al. *"Content-driven, unsupervised clustering of news articles through multiscale graph partitioning."*, Data Science, Journalism & Media workshop (Aug 2018)

1. Unsupervised approach for "hard clustering"
2. Doc2Vec embeddings trained on 5M+ Wikipedia articles
3. Generate a Similarity Graph (Gs) using pairwise cosine similarity between documents in the training set
4. Create a Min Spanning Tree with MST-kNN
5. Markov Stability used to define hard clusters based on partitions extracted using the MST

GOAL 3:
Topic analysis extraction

⇒ Really useful for automated clustering for given no. of topics
⇒ Bit of an overkill for our project

1. We used pre-trained Word2Vec embeddings available with spaCy (en_core_web_md, en_core_web_sm) - each with 300 dimensions

2. We work with the full dataset, not filtering out anything since Topic Modeling is inherently exploratory; additional runs ongoing with notes of length 100+ characters

3. All conflict notes are tokenized by lemmatizing the words, ignoring the standard stopwords, and removing any pronouns after POS-tagging

4. First set of runs are essentially Bag-of-Words with unigrams; second set of runs (ongoing) on bi-grams

GOAL 3:
Topic analysis extraction

- ➔ Goals
- ➔ Recent Literature
- ➔ Pre-processing
- ➔ LDA
- ➔ Visualizations
- ➔ Ongoing/future work

1. We used Latent Dirichlet Allocation (LDA) for the Topic Modeling → most common, easiest to implement method for ‘soft clustering’

2. LDA is very similar to a Bag-of-Words Classifier, but works with inherent frequency distributions instead of optimizing against a target classification

3. Based on a given ‘number of topics’ every conflict note is mapped to each topic through a vocabulary length ‘items’. A sparse matrix (509,157 X 39,799) is then reduced to a denser matrix to assign topic distribution for each conflict note → kind of like training an ‘embedding’ in itself?

GOAL 3: Topic analysis extraction

⇒ Unsupervised approach → no way to assess the model
⇒ Manual review of topics required after each full run

1. **3 Topics (count)** → spontaneity & participation
   Civilian Unrest | Two-sided clash | Single Attack

2. **5 Topics (tfidf)** → strategized & ??
   a. Civilian Unrest / Protest
   b. Civil War Attack
   c. Quantified Airstrikes / Raids
   d. Terrorist Attack
   e. ??

3. **7 Topics (tfidf)**
   → topics are well-defined but dimensions are not
   → geographical factors visible

⇒ Increasing iterations → no significant impact
⇒ Raw Count & TFIDF give vastly different topic extracts

GOAL 3:
Topic analysis extraction

- ➔ Goals
- ➔ Recent Literature
- ➔ Pre-processing
- ➔ LDA
- ➔ Visualizations
- ➔ Ongoing/future work

1. LDA provides p(Topic | Conflict Note) → no objective way to assign one topic for each document

2. We would like to analyze the geographical distribution of Topics & year-on-year trends

3. Compare against ‘event type’ & ‘sub-event type’ provided by the researchers, as also against the ‘interaction’ labels & Aggresor ←→ Victim relations extracted

GOAL 3: Topic analysis extraction

⇒ Simultaneously we want to switch to Continuous BoW
⇒ Using Custom Word Embeddings from Relation Extraction