A NOVEL TRILINGUAL DATASET FOR CRISIS NEWS CATEGORIZATION ACROSS LANGUAGES

Kaisla Kajava

May 27 2021

BACK STORY

- Original idea to use GDELT (Global Database of Events, Language, and Tone) project (Leetaru & Schrodt, 2013):
  - Do some types of conflict or crisis events receive more media attention in some languages and less in others, and
  - Are pre-trained multilingual sentence embeddings useful for clustering news article titles based on the type of conflict or crisis event those articles report?
  - Caveats: non-random distribution, copyright restrictions
- Use Common Crawl instead
  - "the goal of democratizing access to web information by producing and maintaining an open repository of web crawl data that is universally accessible and analyzable" (https://commoncrawl.org/about/, accessed June 1 2021)
  - Create a resource of crisis news articles
  - Cluster crisis news by topic
    - Use pre-trained embeddings
    - Experiment with fine-tuning

MOTIVATION

- Journalists and news agencies monitor vast amounts of streaming data from multiple sources and languages, often in real time
- Actors engaged in crisis management follow news reporting on crises
  - Sellnow et al. (2019): narrative space
  - Baum & Zhukov (2018): co-ownership, market incentives, press freedom
- Pre-trained sentence embedding models enable fast generation of sentence embeddings without costly training
- Multilingual and language-specific models reduce the need for translation pipelines
- No previous research on applying sentence embeddings for clustering crisis news in multiple languages was found

PREVIOUS WORK

- Word2vec (Mikolov et al., 2013), GloVE (Pennington et al., 2014), sent2vec (Pagliardini et al., 2018), doc2vec (Le and Mikolov, 2014) have gained popularity in universal embeddings for short text representation
- Deep pre-trained language models have achieved state-of-the-art results in many NLP downstream tasks
  - Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and its derivative RoBERTa (Liu et al., 2019) provide pre-trained monolingual and multilingual models trained on general-domain corpora that can be fine-tuned
- Most pre-trained language models need annotated training data (fine-tuning)
- Sentence-BERT (Reimers & Gurevych, 2019, 2020) and LASER (Artetxe & Schwenk, 2019) created for improved short text encoding in similarity tasks
- CamemBERT (Martin et al., 2020) trained on the OSCAR corpus (Ortiz Suárez et al., 2020) for French
- SlavicBERT (Arkhipov et al., 2019): BERT-based model trained on Russian news and several Slavic language Wikipedias
- mBERT (Devlin et al., 2018): multilingual BERT model trained on the 102 languages with the largest Wikipedias

PREVIOUS WORK

- CrisisBERT (Liu et al., 2020): trained on crisis-related tweets for crisis event detection on social media
- NewsBERT (Wu et al., 2021): trained for news recommendation and retrieval
- Huang et al. 2020: fine-tuned a pre-trained BERT model in an unsupervised way for text clustering with *k*-means
  - Simultaneously learn text representations and cluster assignments through joint optimization of both the masked language model loss and the clustering-oriented loss, evaluated on question, sentiment, and topic datasets

DATA COLLECTION

- Common Crawl -> comcrawl dumps: domains of selected English-, French-, and Russian-language news sources
- Keyword sets used to filter the news articles into categories: armed conflict/violence, environmental, health, financial
- Example keywords for English:
  - armed conflict/violence: bomb*, missile* , terroris*
  - environmental: pollution, earthquake*, climate change
  - health: epidemi*, vaccin*, covid*
  - financial: subprime, bankrupt*, recession
- Partial manual revision
- These four crisis categories treated as ground truth labels
- Resulting corpus has a total of 38,875 news article entries of various lengths published between 2000-2021
- English data has a total of 11,619,255 word tokens, French 13,310,950, and Russian 10,790,062

DATA DISTRIBUTION

	Armed conflict	Environmental	Health	Financial	Total
English	3,896	1,392	1,726	1,784	8,798
French	8,549	1,193	791	3,806	14,339
Russian	7,601	2,035	2,686	3,416	15,738
Total	20,046	4,620	5,203	9,006	38,875

Table 1: The distribution of unique data points per language and news category.

	Armed conflict	Environmental	Health	Financial	Total
BBC	236	171	51	558	1,016
Daily Mail	1,750	427	280	562	3,019
New York Times	380	190	931	190	1,691
Washington Post	1,530	604	464	474	3,072
Le Figaro	4,421	372	261	1,674	6,728
France 24	1,911	198	345	640	3,094
Le Monde	2,217	623	185	1,492	4,517
Izvestia	4,049	1,681	827	3,320	9,877
RT	3,552	354	1,859	96	5,861

Table 2: The distribution of unique data points per news source and news category.

DATA PREPROCESSING & ISSUES

- For BERT, Sentence-Transformer, and LASER models:
  - names of news agencies omitted
  - stopword removal, lemmatization, and lowercasing were tested but had no effect
- For tf-idf vectorization:
  - lowercasing, lemmatization, and stopword removal had a beneficial effect
  - vectors were reduced for dimensionality using Latent Semantic Analysis
- Overall, some text cleanup such as removing dashes and quotations marks was also done
- Quality differences across languages
- Ambiguous labeling
- Varying lengths of text

PRE-TRAINED SENTENCE EMBEDDING MODELS

- Sentence-Transformer models: `bert-base-nli-cls-token` for English, `paraphrase-xlm-r-multilingual` for French and Russian
- LASER embeddings acquired from pip package
- For comparison, the sentence embedding models were used to encode both full text news articles and article titles, as separate experiments
- Full text articles were assumed to benefit from document-level embedding techniques, so sentence-based models were expected to create better representations for titles
- Clustering was performed using *k*-means with *k*=4 to reflect the number of unique news categories in the dataset
- Different *k* values yielded no significant improvement

FINE-TUNED MODELS

- Fine-tuned on news titles with 5-fold cross-validation
- SlavicBERT for Russian
- CamemBERT for French
- BERT base uncased for English
- mBERT uncased for all
- Adam optimizer learning rate 2e-5, batch size of 32, and sequence length of 128 (sequence length unnecessarily long, since there were less than 30 titles longer than that in the data)

EVALUATION METRICS

- Unsupervised:
  - Silhouette score $SC = \max_k \tilde{s}\ (k)$
  - Adjusted Rand Index $ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}$
  - Adjusted Mutual Information $AMI(U,V) = \frac{MI(U,V) - E\{MI(U,V)\}}{\max\{H(U), H(V)\} - E\{MI(U,V)\}}.$
- Supervised:
  - F1-score $F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$
  - Accuracy

UNSUPERVISED CLUSTERING RESULTS

	AMI	ARI	NMI	Silhouette
EN: `bert-base-nli-cls-token` + *k*-means	.063	.059	.063	.051
EN: tf-idf + *k*-means	.102	.087	.102	.522
FR: `paraphrase-xlm-r-multilingual-v1` + *k*-means	.120	.100	.120	.037
FR: tf-idf + *k*-means	.205	.156	.205	.570
RU: `paraphrase-xlm-r-multilingual-v1` + *k*-means	**.244**	**.166**	**.244**	.042
RU: tf-idf + *k*-means	.197	.070	.197	**.659**

Table 3: The Adjusted Mutual Information, Adjusted Rand Index, Normalized Mutual Information, and Silhouette scores for cluster results on the full text articles when *k*=4. The highest score for each metric is in bold.

	AMI	ARI	NMI	Silhouette
EN: `bert-base-nli-cls-token` + *k*-means	.017	.025	.017	.043
EN: LASER + *k*-means	.048	.034	.048	.032
EN: tf-idf + *k*-means	.047	.070	.052	**.654**
FR: `paraphrase-xlm-r-multilingual-v1` + *k*-means	.075	.028	.075	.030
FR: LASER + *k*-means	.039	.049	.039	.018
FR: tf-idf + *k*-means	.046	.072	.046	.647
RU: `paraphrase-xlm-r-multilingual-v1` + *k*-means	**.125**	.071	**.125**	.046
RU: LASER + *k*-means	.118	**.094**	.118	.118
RU: tf-idf + *k*-means	.032	.040	.032	.582

Table 4: The Adjusted Mutual Information, Adjusted Rand Index, Normalized Mutual Information, and Silhouette scores for cluster results on the article titles when *k*=4. The highest score for each metric is in bold.

CLUSTERING RESULTS ON "GROUND TRUTH"

Cluster	**Precision**	**Recall**	**f1**
conflict	**0.71**	0.46	**0.56**
environmental	0.17	0.34	0.23
health	0.31	**0.53**	0.39
financial	0.70	0.45	0.54

Table 5: The precision, recall, and f1 scores for cluster results on the Russian-language titles when k=4. The highest score for each metric is in bold.

Clusters 0 (conflict) and 3 (financial) had the highest metrics scores, and included words such as: *coronavirus*, *infection*, and *covid*, and *usa*, *trump*, and *election*, respectively

SUPERVISED FINE-TUNING RESULTS

	Accuracy	**Macro f1**
BERT base uncased	0.66	0.63
SlavicBERT	**0.69**	**0.66**
CamemBERT	0.60	0.58
mBERT	0.59	0.56

Table 6: The accuracy and macro f1 scores for each classification experiment on news titles. The highest score for each metric is in bold.

FUTURE WORK

- Manual labeling -> new labels or more data?
- Publication/curation and copyright

REFERENCES

- Arkhipov, M., Trofimova, M., Kuratov, Y. & A. Sorokin (2019). Tuning multilingual transformers for language-specific named entity recognition. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 89–93, Florence, Italy. Association for Computational Linguistics.
- Artetxe, M., & Holger Schwenk (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Trans. Assoc. Comput. Linguistics, 7:597–610
- Baum, M. A. & Y. M. Zhukov (2018). Media Ownership and News Coverage of International Conflict. Political Communication, Volume: 36, issue 1, pages 36-63.
- Devlin, J., Chang, M.-W., Lee, K., & K. Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Computing Research Repository, arXiv:1810.04805.
- Huang, S., Wei, F., Cui, L., Zhang, X. & M. Zhou (2020). Unsupervised Fine-tuning for Text Clustering. Proceedings of the 28th International Conference on Computational Linguistics, pages 5530-5534, https://www.aclweb.org/anthology/2020.coling-main.482.
- Le, Q. V. & T. Mikolov (2014). Distributed Representations of Sentences and Documents. Computing Research Repository, arXiv:1405.4053.
- Leetaru, K. & P. A. Schrodt (2013). GDELT: Global data on events, location, and tone (2013). ISA Annual Convention.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & V. Stoyanov (2019). Roberta: A Robustly Optimized BERT Pre-training Approach. Computing Research Repository, arXiv:1907.11692

REFERENCES

- Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., De la Clergerie, E., Seddah, D., & B. Sagot (2020). CamemBERT: a tasty French language model. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7203–7219, Online. Association for Computational Linguistics.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & J. Dean (2013). Distributed Representations of Words and Phrases and their Compositionality. *Computing Research Repository*, https://arxiv.org/abs/1310.4546.
- Ortiz Suárez, P. J., Romary, L., & B. Sagot (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1703–1714, Online. Association for Computational Linguistics
- Pagliardini, M., Gupta, P. & M. Jaggi (2018). Unsupervised learning of sentence embeddings using compositional n-gram features. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 528–540, New Orleans, Louisiana. Association for Computational Linguistics.
- Pennington, J., Socher, R. & C. Manning (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Reimers, N. & I. Gurevych (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Computing Research Repository, arXiv:1908.10084.344.
- Sellnow, T., Sellnow, D., Helsel, E., Martin, J. & J. Parker (2018). Risk and crisis communication narratives in response to rapidly emerging diseases. *Journal of Risk Research*. 22. 1-12. https://doi.org/10.1080/13669877.2017.1422787