SPICED: NEWS SIMILARITY DETECTION DATASET WITH MULTIPLE TOPICS AND COMPLEXITY LEVELS

- ELENA SHUSHKEVICH (TUD, DUBLIN)
- MANUEL V. LOUREIRO (HUAWEI IRELAND RESEARCH CENTRE)
- LONG MAI (UNIVERSITY COLLEGE DUBLIN)
- STEVEN DERBY (HUAWEI IRELAND RESEARCH CENTRE)
- TRI KURNIAWAN WIJAYA (HUAWEI IRELAND RESEARCH CENTRE)

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

SPICED DATASET

AVAILABLE AT: HTTPS://ZENODO.ORG/RECORD/8044777

- We provide an original dataset of 977 similar news pairs in English (1,954 news articles), devoted to the seven different popular news topics:
    - **C**rime & Law,
    - **C**ulture & Entertainment,
    - **D**isasters & Accidents,
    - **E**conomy & Business,
    - **P**olitics & Conflicts,
    - **S**cience & Technology,
    - **S**ports.

DATASET CREATION AND ANNOTATION

Collecting News Articles:

- **WikiNews**, a Wikimedia Foundation project, follows guidelines requiring topic categorization and support from two independent sources (minimum) with valid URLs.
- Utilized **BeautifulSoup** for web scraping WikiNews articles.
- Focused on **7 categories**.

Measuring Similar News:

- **SimHash** algorithm employed to identify pairs of highly similar articles.
- Validation process ensures pairs originate from the same **WikiNews** webpage.
- **SBERT** utilized to identify the most similar articles within the dataset.

The rules to identify similar news:

- Both news articles in a pair must be about **the same topic and event**;
- Both news articles should have **similar lengths** to avoid information asymmetry;
- **Opinion articles**, prone to biases, should be excluded from similar news classifications;
- Any **numerical values** cited in the articles should be consistent;
- The **time of publication** must be close.

The last step of the filtering:

- **delete duplicate pairs**, which can appear in cases when news articles are devoted to several topics at once.

Topics	CL	CE	DA	EB	PC	ST	SP
Filters							
SimHash	76,996	8,672	24,015	30,291	123,791	8,916	14,954
Source of the same Wikinews page	511	259	316	312	822	273	334
SBERT	501	230	300	279	779	249	318
Experts' annotation	238	95	137	120	361	136	94
Duplicates removal	192	90	124	107	259	111	94

COMPLEXITY LEVELS

Inter-Topic:

- Positive samples: Similar news pairs.
- Negative samples: Dissimilar news pairs from different topics.

Intra-Topic:

- Positive and negative pairs within the same topic.
- Seven subsets corresponding to different topics.
- Exclusion of challenging examples from negative pairs.

Hard Examples:

- Positive pairs and 3,000 most similar negative pairs within each intra-topic.

Combined:

- Union of positive and negative news pairs from previous sets.

Model	MinHash	BERT	SBERT	SimCSE
Inter-Topic				
F1-score	*0.707*	*0.786*	*0.920*	*0.896*
Intra-Topic				
Crime & Law (CL)	0.816	0.851	0.957	0.957
Culture & Entertainment (CE)	0.902	0.923	0.923	0.943
Disaster & Accidents (DA)	0.742	0.853	0.935	0.853
Economy & Business (EB)	0.678	0.828	0.899	0.937
Politics & Conflict (PC)	0.650	0.776	0.911	0.875
Science & Technology (ST)	0.690	0.824	0.921	0.847
Sporting Activities (SP)	0.840	0.840	0.982	0.816
Average F1-score	*0.760*	*0.842*	*0.933*	*0.890*
Hard Examples				
Crime & Law (CL)	0.727	0.891	0.935	0.919
Cultu921403-re & Entertainment (CE)	0.833	0.906	0.902	0.943
Disaster & Accidents (DA)	0.742	0.795	0.938	0.868
Economy & Business (EB)	0.690	0.774	0.952	0.909
Politics & Conflict (PC)	0.702	0.829	0.940	0.892
Science & Technology (ST)	0.741	0.639	0.853	0.667
Sporting Activities (SP)	0.840	0.840	0.945	0.964
Average F1-score	*0.754*	*0.811*	*0.924*	*0.880*
Combined				
F1-score	*0.757*	*0.799*	*0.922*	*0.875*

Conclusions:

- Proposed a **novel semantic textual similarity dataset** for news data, considering emergent semantic categories.
- Created **32 training and test datasets** for news similarity detection, organized into four approaches: **Inter-Topic**, **Intra-Topic**, **Hard Example** Mining, and **Combined Similarity**.
- **Experimental results** highlight the challenge posed by our dataset for state-of-the-art models (MinHash demonstrated the lowest results, SBERT - the highest results).