Data Model Canvas - SCOUT

Business Problem/Question

1. Lack of easily and quickly accesible information about global response (per country) on certain issues/events based on local media.
2. Linguistic barriers for general understanding of foreign articles.

Scientific Value

1. Sentiment analysis tools for languages other than English.
2. Issue clustering for varying languages.
3. Clean data set on sentiment response and issue clustering.

Business Value

1. Aggregated information on global ambience in certain areas for conscious decision making.

Team/Collaborations

Data Crawling
Data Processing
NLP – Sentiment Analysis
NLP – Issue Clustering
Deep Learning
Basic Machine Learning
Language knowledge
Web Application Development

Partners/Collaborations

- Prof. dr hab. inż. **Przemysław Kazienko**
- **USA** – Younique Bales and Ted Mach
- **UK** – Karim Amadi
- **Spain** - Vanesa Jiménez Molina
- **Slovenia** - Institut Jožef Stefan

Data

1. Crawled and manually labeled data.
2. Event Registry articles.
3. The GDELT Project (especially for issue clustering).

Model

1. Data aggregation from various sources
2. Data preprocessing.
3. Sentiment Analysis model
4. Issue Clustering model
   a) Possibly Named Entity Recognition
   b) Model/Module for identifying related issues across languages
5. Possibly Google Translate API usage for support between modules.

The model assumes heavy use of transfer learning from language models trained on Wikipedia (**Language Model Zoo**), as proposed in **Universal Language Model Fine-Tuning** for Text Classification.

Evaluation

1. Model-specific evaluation metrics (for sentiment analysis against ground truth).
2. Benchmark data sets for Polish, Spanish and English (IMDb, PolEval, ISOL).
3. For highly aggregated results confrontation against local knowledge (partners).

Deployment/UX

1. Public result presentation to all engaged in the project and anyone interested.
2. Possibly a publication.
3. Web application for convinient browsing and visualizing achieved results.

Users

1. People interested in global-scale view on media.
2. Data Scientists
3. Sociologists
4. PR-related field specialists
5. Journalists

Expected Costs

1. Application server and domain costs.
2. Final presentation organisation costs.
3. Application Maintenance.
4. Cloud or local services for computing, power consumption.
5. Possible conference costs.

Expected Benefits

1. Tool for obtaining quick information on current issues in certain countries and their perception.
2. Data set for usage by other scientists.
3. Possibly scientific paper on one of the components.