Detecting Emerging Technologies

In the paper, authors have presented a multi-layer quantitative approach able to identify future signs from scientific publications on different technical terms by leveraging deep learning and weak signal analysis. First, they improve upon keyword extraction by going beyond traditional statistical approaches and leverage a pre-trained transformer-based deep learning technique, namely BERT. Second, they apply recent concepts of weak signal analysis to the early detection of technology emergence. Weak signal is defined as a signal that can affect the business or its environment, could be difficult to track, and needs time to become mature and mainstream. Weak signals can be regarded as signals about future trends of a given technology

The implementation starts with splitting the data into separate data intervals. Authors divided the examined time period into 3-year time intervals, starting from the [1985, 1987] period to [2018, 2020]. They referred to these intervals as t1 to t12. Pre-trained BERT model was applied for keyword extraction in each of these time intervals. The keywords of each time interval were combined and duplicates were removed.

Keyword Emergence Map

Term frequencies were used to calculate keywords' degree of visibility. A time-weighted formula is used to calculate the degree of visibility (DoV) of the extracted keywords in each of the 12 aforementioned time intervals, i.e., t1 to t12:

$$DoV_{ij} = \left(\frac{TF_{ij}}{N_j}\right) * \left(1 - w * (n - j)\right)$$

where TF_{ij} is the total frequency of keyword i in time interval j, N_j is the total number of documents in time interval j, n is the number of time intervals (n = 12, in our experiment), and w (=0.05) is a constant time-weight.

The estimated DoV values can then be used to generate the Keyword Emergence Map (KEM). In KEM, the x-axis shows the average term frequencies of the keywords, while the y-axis shows the geometric mean-estimated growth rate of the DoV. The medians of the corresponding values on each axis are used to divide the KEM's quadrants. In the KEM map, this results in the 4 quadrants (using median values): *Strong signals* are represented in the top-right region by keywords with a high average term frequency and have a high average DoV growth rate; *weak signals* are represented in the top-left region by keywords with a low average term frequency and a high average DoV growth rate; *latent signals* are represented by keywords in the bottom-left region and bottom right of the KEM represents *well-known but not strong signals* that are keywords with high term frequency but low DoV growth rate.

Keyword Issue Map

Following time-weighted formula is used to calculate the degree of diffusion (DoD) of the extracted keywords in each of the 12 aforementioned time intervals, i.e., t1 to t12:

$$DoD_{ij} = \left(\frac{DF_{ij}}{N_j}\right) * (1 - w * (n - j))$$

DF_{ij} is the document frequency of keyword i in time interval j.

The average time-weighted growing rate of keyword document frequencies was plotted on the y-axis, and the average keyword document frequencies were plotted on the x-axis, to create the Keyword Issue Map (KIM). Weak signals relate to terms with a strong growth rate but a low document frequency.

Signals appearing in the same quadrant in both KEM and KIM indicated similar degrees of visibility and dissemination and were extracted. Signals in each quadrant had the same interpretation as mentioned in the KEM.

Results:

Neural Information Processing Systems (NIPS) is one of the top machine learning conferences in the world. It covers topics ranging from deep learning and computer vision to cognitive science and reinforcement learning. We have used the NIPS papers ranging from the first 1987 conference to the current 2016 conference for the validation of our implementation.

We have split the data into 11 time intervals, 3 years in the first 10 intervals and 2 years in the last interval. We have extracted 30 monogramic keywords from each of the time intervals using the default KeyBert model. After removal of duplicates and stop-words, we were left with 276 unique keywords. KEM and KIM matrices are computed using the DoV and Dod values respectively. After taking the intersection of words in each quadrant, we got the following results. (few words from each category)

1. **Strong signal**: ['classify', 'auxiliary', 'probabilistic', 'physiological', 'search', 'automated', 'integration', 'routine', 'kernel', 'transductive', 'noising', 'learns', 'multivariate', 'mcmc', 'multitask', 'extraction', 'blockmodel', 'decoder', 'integrate', 'variational', 'column', 'mixture', 'reliability', 'manual', 'backpropagation', 'orientation', 'quantization', 'memory', 'analytic', 'datasets', 'grouping', 'mse', 'programming', 'frequency', 'discus', 'gate', 'hmms', 'ising', 'modelling', 'discretization', 'optimum', 'training', 'navigation', 'lstm', 'serve', 'dynamical', 'select', 'tive', 'mle', 'semidefinite', 'bipartite', 'characteristic', 'dynamic', 'filtering', 'forecasting', 'likelihood', 'copula', 'rnns', 'devise', 'ranking', 'dataset', 'multilayer',

'eigenvectors', 'ridge', 'handwritten', 'modeling', 'sensor', 'recognition', 'nonlinearities', 'hessian', 'empirical', 'convergence', 'perceptrons', 'backprop', 'minimizer', 'require', 'cognitive', 'iterative', 'patterns', 'protein', 'benchmarks', 'dynamically', 'ica', 'especially', 'rnn', 'learnable', 'controller', 'distance', 'boost', 'decoding', 'np', 'embeddings', 'architectures', 'selects', 'hierarchical', 'identifiability', 'encoder', 'iteration', 'specialized', 'tuning', 'functions', 'gaussian', 'classifiers', 'parameterize', 'discount', 'robot', 'svms', 'hebb', 'average', 'activation', 'leveraging', 'rotation', 'integrates', 'neuron', 'classifier', 'incremental', 'batch', 'multimodal', 'reinforcement', 'cluster', 'autoencoders', 'multiscale']

2. **Weak signa**l: ['i100', 'kanji', 'requirements', 'gesture', 'tensorflow']
3. **Not Strong But Well Known**: ['retinal', 'anytime', 'synapse', 'classifying', 'markovian']
4. **Latent Signal**: ['gaussianization', 'kbsvms', 'neurocomputer', 'zype', 'infant', 'superpixels', '1993', 'sde', 'exif', 'long_algorithms_2013', 'memorization', 'rankprop', 'stepwise', 'autoassociator', 'datamatrix', 'wavelets', 'speechreading', 'locomotory', 'multisite', 'manipulator', 'cochlear', 'dialysis', 'electroencephalographic', 'unfolding', 'emulation', 'gestures', 'anna', 'subsampled', '21578', 'muscle', 'helicopter', 'optoelectronic', 'synergistic', 'mainlifting', 'emulates', 'holistic', 'footnote', 'factorizing', 'audiometric', 'jain', 'neurophysiological', 'reconfigurable', '10184', 'accumulator', 'letterforms', 'approximability', 'lqg', 'spectrogram', 'abstracts', 'aibo', 'reconfigurability', 'kolmogorov', 'hill', 'millisecond', 'pulsestream', 'echolocating', 'benchmarking', 'phonetic', 'supermodularity', 'factorie', 'piezometer', 'sparsemax', 'simplemkl', 'wmt', 'berthet2013computational', 'eigenmaps', 'logconcave', 'subgrouping', 'blockmodels', 'matlab', 'sexnet', 'ssth', 'recnorm', 'nagumo', 'fractal', 'electroencephalogram', 'sodium', 'sigmoidal', 'matchingexpectations', 'batra', 'percentile', 'thud', 'resnet152', 'controllable', 'optimizers', 'metabolomics', 'topdown', 'recipes', 'microelectronic', 'impaired', 'fitc', 'songs', 'badanidiyuru', 'ucsd', 'knapsack', 'elegans', 'automaton', 'logitboost', 'gacv', 'connectomics', 'electrotonic', 'inforamtion', 'cvp', 'arrhythmia', 'psoms', 'ecg', 'algorithmthat', 'dataflow', 'memorize', 'movielens', 'splining', 'tokamak', 'agrees', 'interneurons', 'neuromodulation']

Results Description

1. Weak signals are early signs of possible but not confirmed changes that may later become more significant indicators of critical forces. They can be regarded as signals about future trends of a given technology.
2. Strong signals contain keywords which were the state-of-the-art technology of a particular period. Weak signals have the potential to become strong signals in the near future.

3. Latent Signals contain keywords which are not well known and less likely to emerge in the near future.
4. Not strong but well known signal refers to some common terms which are in general use.