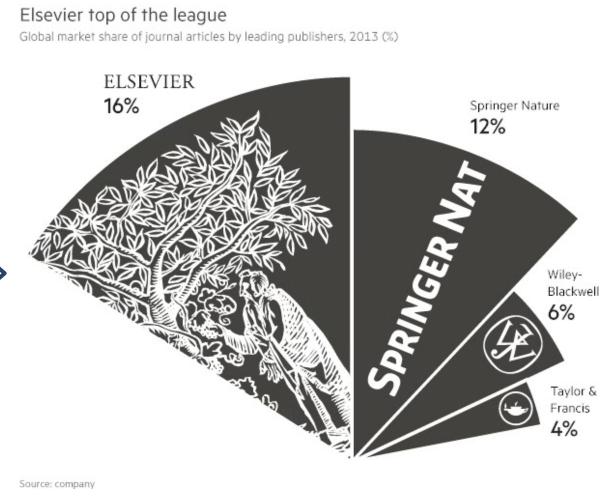
CNRS - Toulouse INP - UT3 - UT Capitole - UT2

Institut de Recherche en Informatique de Toulouse

Mining tortured acronyms from the scientific literature

Alexandre Clausse, Guillaume Cabanac, Pascal Cuxac, Cyril Labbé

OP14.3



https://www.ft.com/content/93138f3e-87d6-11e5-90de-f44762bf9896

voice recognition

deep neural network

Convolutional Neural Network. Artificial Neural Networks with numerous layers are termed as Deep Neural Networks or Deep Learning. It has been explored as one another key resource in recent years and has become quite well recognized in the literary community because of its efficiency to manage with huge amounts of data [17]. The most well-known profound neural network is the Convolutional Neural Networks (CNNs), which takes its name from operation of mathematical dimension from the matrixes termed convolution. Convolutional Neural Network (CNN) has various types of layers; it includes pooling, nonlinearity, and convolutional and fully connected layers. Convolutional Neural Network has pivotal outcomes over previous decades in an assortment of fields identified with design acknowledgment, from picture handling to voice acknowledgment [8]. The significant part of CNN is to get theoretical highlights when information proliferates towards the more profound layers. For instance, in picture characterization, the edge may be distinguished in the principal layers, and afterward the less difficult shapes in the subsequent layers, and afterward the more elevated level highlights.

pattern recognition

deep layers

high-level features

pubpeer.com/publications/99826B46D2BA62C9FA87CAF11AE0FD

References: [3] & [4]

spectral local linearization method

All Images Videos Shopping Forums ⋮ More

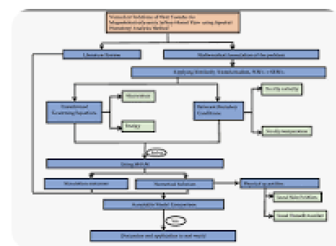
Scholarly articles for **spectral local linearization method**

A new **spectral local linearization method** for nonlinear … - Motsa - Cited by 173

… conditions using **spectral local linearization method** - Sithole - Cited by 23

Spectral methods to solve nonlinear problems: A … - Rai - Cited by 27

The spectral local linearization method. The unearthly neighborhood linearization technique (SLLM) can be used to unravel an arrangement of the non-comparable conditions. Within this method the criteria for logarithmic criteria are linearized using thoughts similar to the Gauss–Seidel approach.



ScienceDirect.com
https://www.sciencedirect.com › science › article › pii ⋮

Spectral methods to solve nonlinear problems: A review

About featured snippets • Feedback

https://web.archive.org/web/20240521113315/https://www.google.com/search?client=firefox-b-d&q=spectral+local+linearization+method#ip=1

(snapshot taken on May 21, 2024)

spectral local linearization method

All Images Videos Shopping Forums More

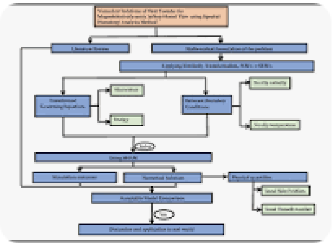
Scholarly articles for spectral local linearization method

A new **spectral local linearization method** for nonlinear … - Motsa - Cited by 173

… conditions using **spectral local linearization method** - Sithole - Cited by 23

Spectral methods to solve nonlinear problems: A … - Rai - Cited by 27

The spectral local linearization method. The unearthly neighborhood linearization technique (SLLM) can be used to unravel an arrangement of the non-comparable conditions. Within this method the criteria for logarithmic criteria are linearized using thoughts similar to the Gauss–Seidel approach.



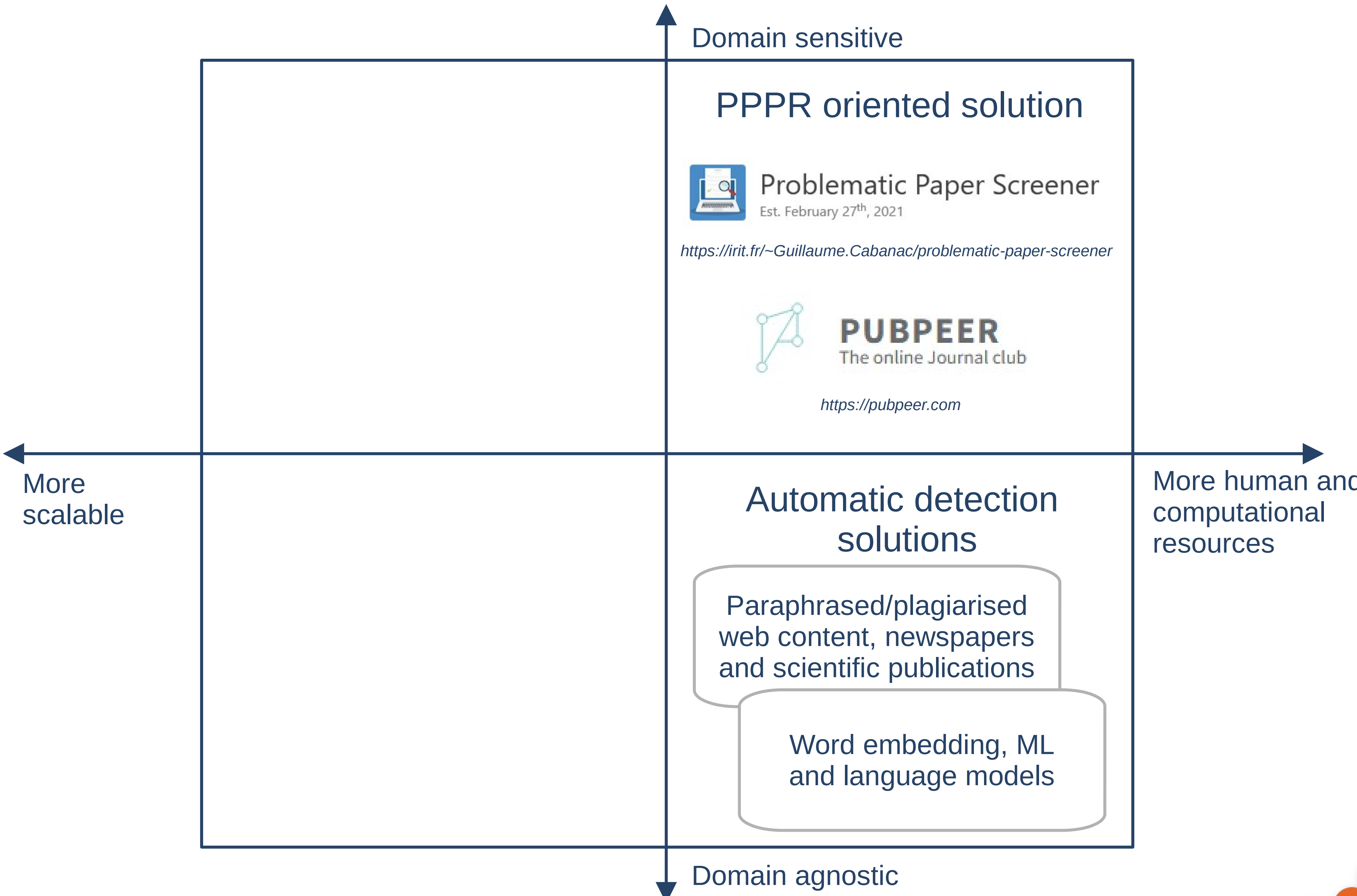
ScienceDirect.com
https://www.sciencedirect.com › science › article › pii

Spectral methods to solve nonlinear problems: A review

About featured snippets • Feedback

https://web.archive.org/web/20240521113315/https://www.google.com/search?client=firefox-b-d&q=spectral+local+linearization+method#ip=1

(snapshot taken on May 21, 2024)



References: [1], [2], [5], [7] to [11]

Domain sensitive

PPPR oriented solution

Problematic Paper Screener
Est. February 27th, 2021

https://irit.fr/~Guillaume.Cabanac/problematic-paper-screener

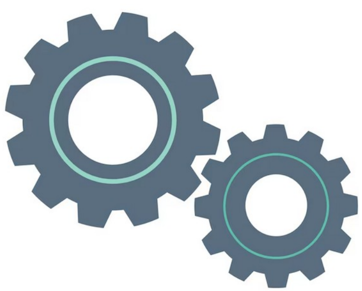
PUBPEER
The online Journal club

https://pubpeer.com

More scalable

More human and computational resources

Document oriented solution



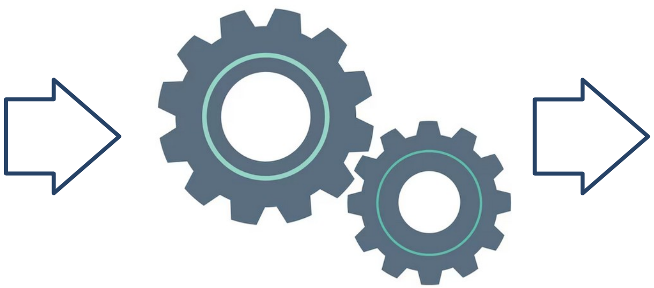
Automatic detection solutions

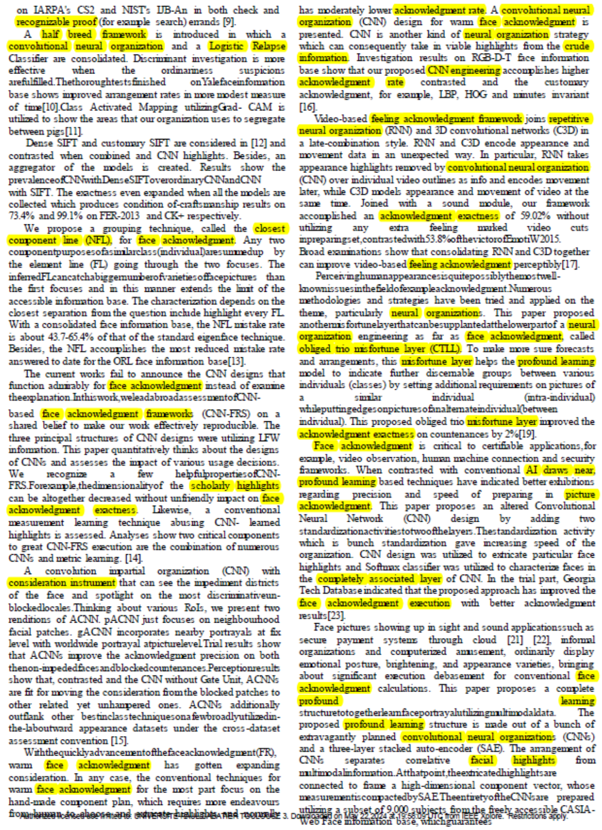
Paraphrased/plagiarised web content, newspapers and scientific publications

Word embedding, ML and language models

Domain agnostic

References: [1], [2], [5], [7] to [11]





on IARPA's CS2 and NIST's IJB-An in both check and recognizable proof (for example search) errands [9].

A half breed framework is introduced in which a convolutional neural organization and a Logistic Relapse Classifier are consolidated. Discriminant investigation is more effective when the ordinariness suspicions arefulfilled.Thethoroughtestsfinished onYalefaceinformation base shows improved arrangement rates in more modest measure of time[10].Class Activated Mapping utilizingGrad- CAM is utilized to show the areas that our organization uses to segregate between pigs[11].

Dense SIFT and customary SIFT are considered in [12] and contrasted when combined and CNN highlights. Besides, an aggregator of the models is created. Results show the prevalenceofCNNwithDenseSIFToverordinaryCNNandCNN with SIFT. The exactness even expanded when all the models are collected which produces condition of-craftsmanship results on 73.4% and 99.1% on FER-2013 and CK+ respectively.

We propose a grouping technique, called the closest component line (NFL), for face acknowledgment. Any two componentpurposesofasimilarclass(individual)aresummedup by the element line (FL) going through the two focuses. The inferredFLcancatchabiggernumberofvarietiesoffacepictures than the first focuses and in this manner extends the limit of the accessible information base. The characterization depends on the closest separation from the question include highlight every FL. With a consolidated face information base, the NFL mistake rate is about 43.7-65.4% of that of the standard eigenface technique. Besides, the NFL accomplishes the most reduced mistake rate answered to date for the ORL face information base[13].

The current works fail to announce the CNN designs that function admirably for face acknowledgment instead of examine theexplanation.Inthiswork,weleadabroadassessmentofCNN-based face acknowledgment frameworks (CNN-FRS) on a shared belief to make our work effectively reproducible. The three principal structures of CNN designs were utilizing LFW information. This paper quantitatively thinks about the designs of CNNs and assesses the impact of various usage decisions. We recognize a few helpfulpropertiesofCNN-FRS.Forexample,thedimensionalityof the scholarly highlights can be altogether decreased without unfriendly impact on face acknowledgment exactness. Likewise, a conventional measurement learning technique abusing CNN- learned highlights is assessed. Analyses show two critical components to great CNN-FRS execution are the combination of numerous CNNs and metric learning. [14].

A convolution impartial organization (CNN) with consideration instrument that can see the impediment districts of the face and spotlight on the most discriminativeun-blockedlocales.Thinking about various RoIs, we present two renditions of ACNN. pACNN just focuses on neighbourhood facial patches. gACNN incorporates nearby portrayals at fix level with worldwide portrayal atpicturelevel.Trial results show that ACNNs improve the acknowledgment precision on both thenon-impededfacesandblockedcountenances.Perceptionresults show that, contrasted and the CNN without Gate Unit, ACNNs are fit for moving the consideration from the blocked patches to other related yet unhampered ones. ACNNs additionally outflank other bestinclasstechniquesonafewbroadlyutilizedin-the-laboutward appearance datasets under the cross-dataset assessment convention [15].

Withthequicklyadvancementofthefaceacknowledgment(FR), warm face acknowledgment has gotten expanding consideration. In any case, the conventional techniques for warm face acknowledgment for the most part focus on the hand-made component plan, which requires more endeavours from human to choose and

has moderately lower acknowledgment rate. A convolutional neural organization (CNN) design for warm face acknowledgment is presented. CNN is another kind of neural organization strategy which can consequently take in viable highlights from the crude information. Investigation results on RGB-D-T face information base show that our proposed CNN engineering accomplishes higher acknowledgment rate contrasted and the customary acknowledgment, for example, LBP, HOG and minutes invariant [16].

Video-based feeling acknowledgment framework joins repetitive neural organization (RNN) and 3D convolutional networks (C3D) in a late-combination style. RNN and C3D encode appearance and movement data in an unexpected way. In particular, RNN takes appearance highlights removed by convolutional neural organization (CNN) over individual video outlines as info and encodes movement later, while C3D models appearance and movement of video at the same time. Joined with a sound module, our framework accomplished an acknowledgment exactness of 59.02% without utilizing any extra feeling marked video cuts inpreparingset,contrastedwith53.8%ofthevictorofEmotiW2015. Broad examinations show that consolidating RNN and C3D together can improve video-based feeling acknowledgment perceptibly[17].

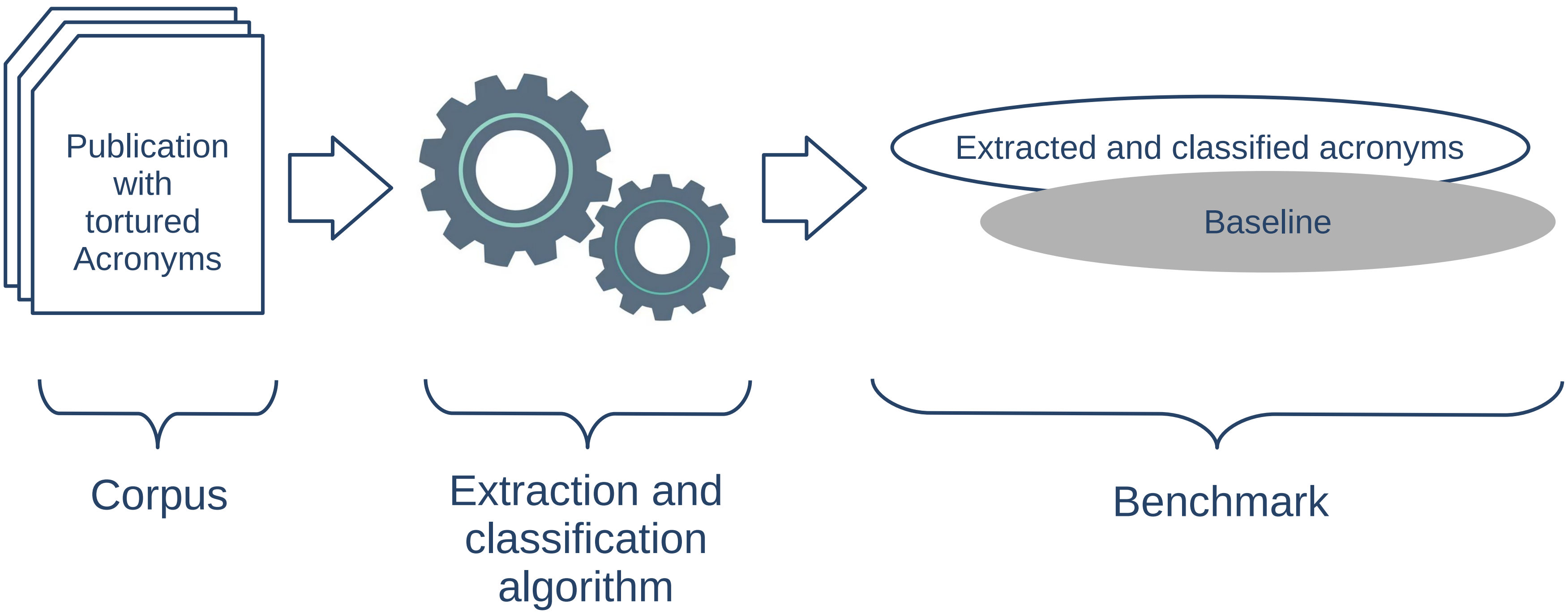
Perceivinghumanappearancesisquitepossiblythemostwell-knownissuesinthefieldofexampleacknowledgment.Numerous methodologies and strategies have been tried and applied on the theme, particularly neural organizations. This paper proposed anothermisfortunelayerthatcanbesupplantedatthelowerpartof a neural organization engineering as far as face acknowledgment, called obliged trio misfortune layer (CTLL). To make more sure forecasts and arrangements, this misfortune layer helps the profound learning model to indicate further discernable groups between various individuals (classes) by setting additional requirements on pictures of a similar individual (intra-individual) whileputtingedgesonpicturesofanalternateindividual(between individual). This proposed obliged trio misfortune layer improved the acknowledgment exactness on countenances by 2%[19].

Face acknowledgment is critical to certifiable applications,for example, video observation, human machine connection and security frameworks. When contrasted with conventional AI draws near, profound learning based techniques have indicated better exhibitions regarding precision and speed of preparing in picture acknowledgment. This paper proposes an altered Convolutional Neural Network (CNN) design by adding two standardizationactivitiestotwoofthelayers.Thestandardization activity which is bunch standardization gave increasing speed of the organization. CNN design was utilized to extricate particular face highlights and Softmax classifier was utilized to characterize faces in the completely associated layer of CNN. In the trial part, Georgia Tech Database indicated that the proposed approach has improved the face acknowledgment execution with better acknowledgment results[23].

Face pictures showing up in sight and sound applicationssuch as secure payment systems through cloud [21] [22], informal organizations and computerized amusement, ordinarily display emotional posture, brightening, and appearance varieties, bringing about significant execution debasement for conventional face acknowledgment calculations. This paper proposes a complete profound learning structuretotogetherlearnfaceportrayalutilizingmultimodaldata. The proposed profound learning structure is made out of a bunch of extravagantly planned convolutional neural organizations (CNNs) and a three-layer stacked auto-encoder (SAE). The arrangement of CNNs separates correlative facial highlights from multimodalinformation.Atthatpoint,theextricatedhighlightsare connected to frame a high-dimensional component vector, whose measurementiscompactedbySAE.TheentiretyoftheCNNsare prepared utilizing a subset of 9,000 subjects from the freely accessible CASIA-Web Face information base, whichguarantees

https://pubpeer.com/publications/2432C53B542C6B29AF449894368D61

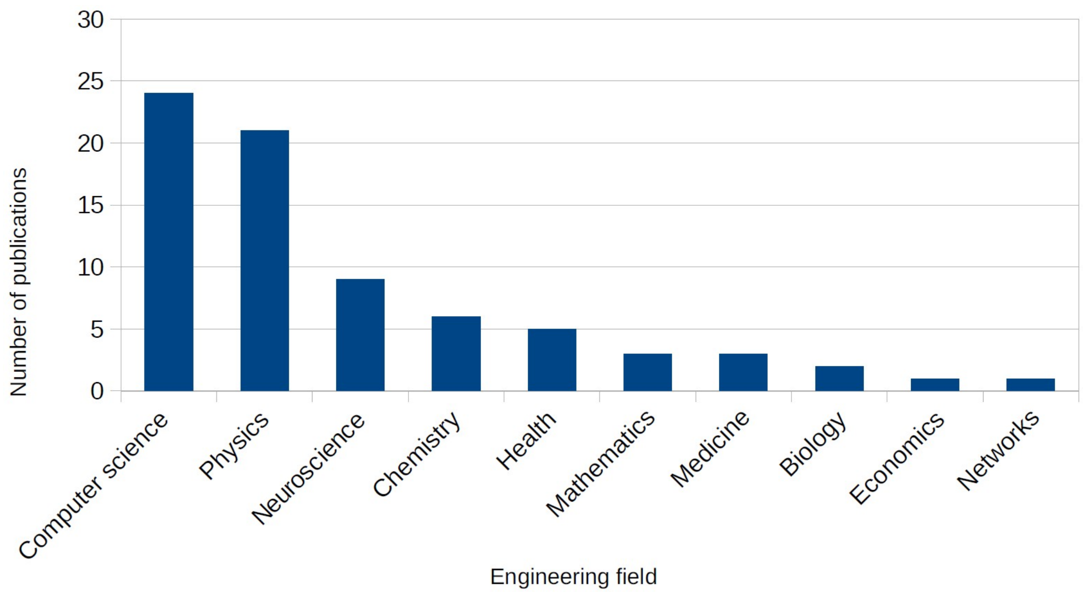
Submission (authors) → **Editorial workflow (publishers)** → **Decision making (editors)**

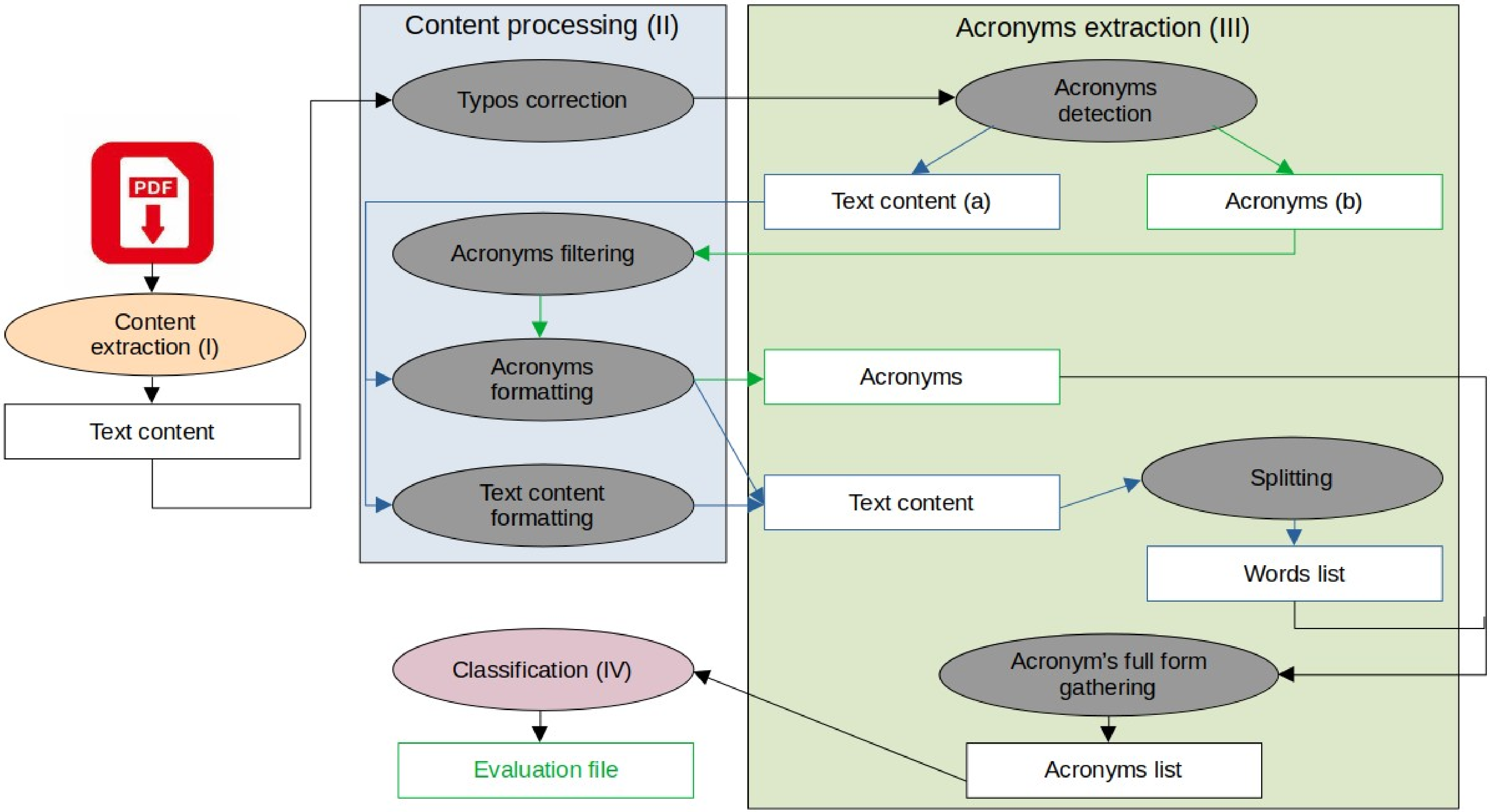


Bolster Vector Machine (SVM) → *Support Vector Machine* (SVM)

- 75 open access publications
- Published between 2015 and 2023
- 37 multidisciplinary publications
- 995 acronyms
- 366 tortured acronyms (36.8%)
- 46 invalid acronyms (4.6%)

http://doi.org/10.5281/zenodo.10492230




Content processing (II)
Acronyms extraction (III)
PDF
Typos correction
Acronyms detection
Text content (a)
Acronyms (b)
Acronyms filtering
Content extraction (I)
Acronyms formatting
Acronyms
Text content
Text content formatting
Text content
Splitting
Words list
Classification (IV)
Acronym's full form gathering
Evaluation file
Acronyms list

Characteristic	Class	
	Genuine	**Tortured**
(1) Same amount of initials	*Hidden Markov Model (HMM)*	*Concealed Markov Display (HMM)*
(2) Different amount of initials	*Generalised Linear Model (GLM)*	*Summed Up Direct Models (GLM)*
(3) Compounds	*Multidrug Resistance (MDR)*	*Multidrug Safe (MDR)*
(3) Chemicals	*Diethylenetriamine Pentaacetic acid (DTPA)*	*Diethylenetriamine Pentaacetic ruinous (DTPA)*
(4) Reduced forms	*Residual neural Network (ResNet)*	*Lingering brain Organization (ResNet)*

Characteristic
(1) Same amount of initials
(2) Different amount of initials
(3) Compounds
(3) Chemicals
(4) Reduced forms

Class

// (1) no offset
using Hidden Markov Model (HMM) → **Genuine**
using Concealed Markov Display (HMM) → Next step

// (2) no offset
~~the~~ Moroccan Agency ~~of~~ Press (MAP) → **Genuine**
~~the~~ Moroccan Organization ~~of~~ Press (MAP) → Next step

// (1) 1 token offset
Hidden Markov Model architecture (HMM) → **Genuine**
Concealed Markov Display architecture (HMM) → Next step

// (2) 1 token offset
Center ~~for~~ Disease Control ~~and~~ Prevention (CDC) → **Genuine**
Center ~~for~~ Illness Control ~~and~~ Prevention (CDC) → Next step

// (3)
[…] Multidrug Resistance […] (MDR) → **Genuine**
[…] Multidrug Safe […] (MDR) → Next step

// (4)
[…] Residual Network […] (ResNet) → **Genuine**
[…] Lingering Organization […] (ResNet) → **Tortured**

- **Recall** (R) : TP / (TP + FN)
- **Precision** (P) : TP / (TP + FP)
- **F-score** (F1) : 2 * (P * R) / (P + R)
- Invalid acronyms are considered as genuine

Acronyms extraction

	Annotation	Prediction
TP	Acronym	Acronym
FP	-	Acronym
TN	Acronym	-
FN	-	-

Acronyms classification

	Annotation	Prediction
TP	Tortured	Tortured
FP	Genuine	Tortured
TN	Genuine	Genuine
FN	Tortured	Genuine

Tortured acronyms extraction

	Annotation	Prediction
TP	Tortured	Tortured
FP	Genuine Genuine -	Tortured - Tortured
TN	Tortured Tortured -	Genuine - Genuine
FN	-	-

Task	Recall	Precision	F1
Extraction	0.99	0.89	0.94
Classification	0.95	0.79	0.86
Tortured acronyms extraction	0.93	0.62	0.74

- Domain agnostic
- Scalable
- PDF-oriented
- No need for fingerprints
- Found 185 new fingerprints for the PPS

- Rework the filtering rules (hallucinated and polysemic expressions)
- From silver to gold standard
- Compute weighted evaluation metrics
- Evaluate the pipeline on a new dataset

The case of IOP Publishing:

- Publications between 2017 and 2023
- 3848 problematic publications
- 10% of problematic publications are tortured
- At least 7% contain tortured acronyms

Source: the PPS (consulted on May 21, 2024)

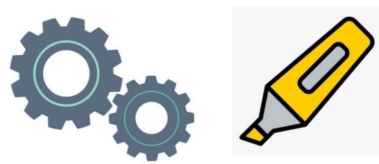
Hallucinated acronym: *Bolster Vector Machine* (BVM) → *Support Vector Machine* (SVM)
Polysemic expression: *Profound Learning*

Draft

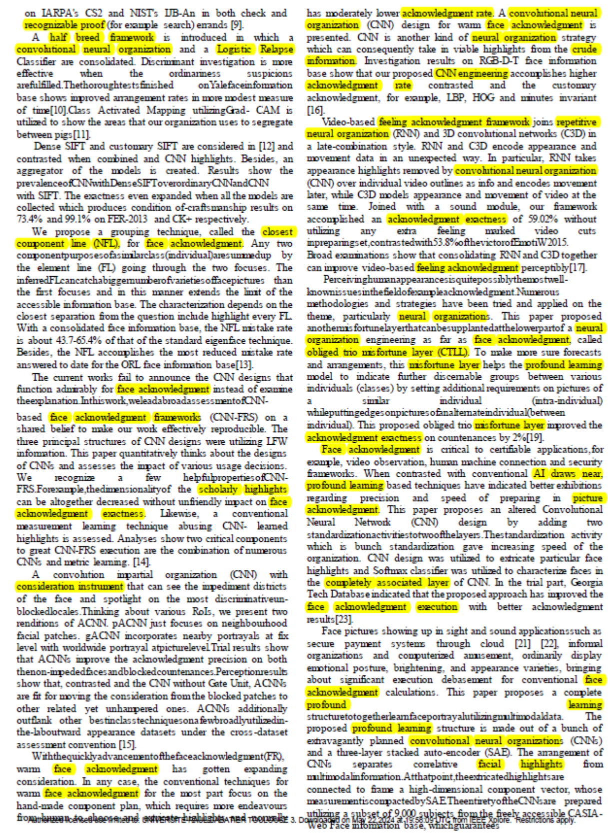
Preprint

Article

Tortured Phrases Tool Kit (TPTK)



Tortured phrases extraction and highlighting



https://pubpeer.com/publications/2432C53B542C6B29AF449894368D61

Submission (authors)

Editorial workflow (publishers)

Decision making (editors)

Bibliography

[1] Barbour B., Stell B. M., “PubPeer: Scientific Assessment Without Metrics”, in M. Biagioli, A. Lippman (ed.), Gaming the Metrics: Misconduct and Manipulation in Academic Research, The MIT Press, p. 149–155, 2020.
DOI : https://doi.org/10.7551/mitpress/11087.003.0015.

[2] Becker J., Wahle J. P., Ruas T., Gipp B., “Paraphrase Detection: Human vs. Machine Content”, Eighth Annual Conference on Advances in Cognitive Systems, Advances in Cognitive Systems, 2020.
URL : https://advancesincognitivesystems.github.io/acs/data/ACS2020_paper_28.pdf.

[3] Biagioli M., Lippman A. (ed.), Gaming the Metrics: Misconduct and Manipulation in Academic Research, The MIT Press, 2020.
DOI : https://doi.org/10.7551/mitpress/11087.001.0001.

[4] Cabanac G., Labbe C., Magazinov A., “Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals”, 2021. Preprint arXiv.
DOI : https://doi.org/10.48550/arXiv.2107.06751.

[5] Cabanac G., Labbe C., Magazinov A., “The ‘Problematic Paper Screener’ automatically selects suspect publications for post-publication (re)assessment”, 2022. Preprint arXiv.
DOI : https://doi.org/10.48550/arXiv.2210.04895.

[6] Clausse A., Cabanac G., Cuxac P., Labbe C., “Extraction d’acronymes tortures dans la litterature scientifique”, in P. Cuxac, C. Lopez (ed.), Atelier TextMine de la conference Extraction et Gestion des Connaissances (EGC) de 2024, Dijon, France, p. 27–37, 2024.
URL : https://hal.science/hal-04426448.

[7] Gehrmann S., Strobelt H., Rush A., “GLTR: Statistical Detection and Visualization of Generated Text”, in M. R. Costa-jussa, E. Alfonseca (ed.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, p. 111–116, 2019.
DOI : https://doi.org/10.18653/v1/P19-3019.

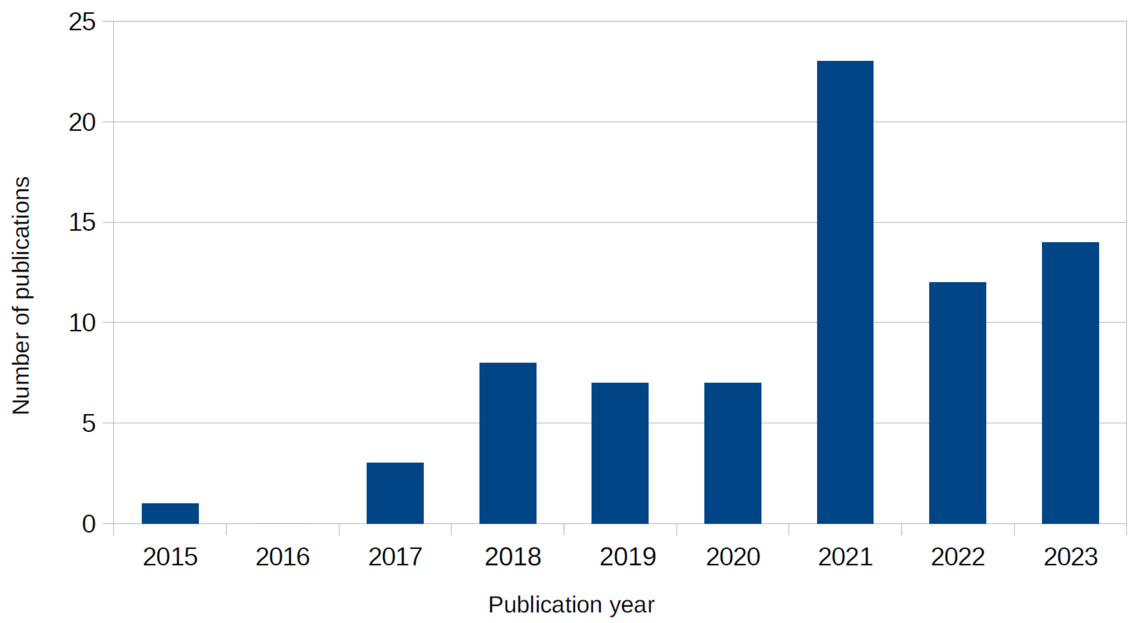
[8] Lay P., Lentschat M., Labbe C., “Investigating the detection of Tortured Phrases in Scientific Literature”, in A. Cohan, G. Feigenblat, D. Freitag, T. Ghosal, D. Herrmannova, P. Knoth, K. Lo, P. Mayr, M. Shmueli-Scheuer, A. de Waard, L. L. Wang (ed.), Proceedings of the Third Workshop on Scholarly Document Processing, Association for Computational Linguistics, p. 32–36, 2022.
URL : https://aclanthology.org/2022.sdp-1.4.

[9] Macbeth J. C., Chang E., Chen J. G., Grandic S., “A Broader Range for ‘Meaning the Same Thing’: Human Against Machine on Hard Paraphrase Detection Tasks”, 2020.
URL : https://api.semanticscholar.org/CorpusID:232247625.

[10] Martel E., Lentschat M., Labbe C., “Detection of Tortured Phrases in Scientific Literature”, in T. Ghosal, F. Grezes, T. Allen, K. Lockhart, A. Accomazzi, S. Blanco-Cuaresma (ed.), Proceedings of the Second Workshop on Information Extraction from Scientific Publications, Association for Computational Linguistics, p. 43–48, 2023.
DOI : https://doi.org/10.18653/v1/2023.wiesp-1.6.

[11] Wahle J. P., Ruas T., Foltynek T., Meuschke N., Gipp B., “Identifying Machine-Paraphrased Plagiarism”, in M. Smits (ed.), Information for a Better World: Shaping the Global Future, Springer, p. 393–413, 2022.
DOI : https://doi.org/10.1007/978-3-030-96957-8_34.

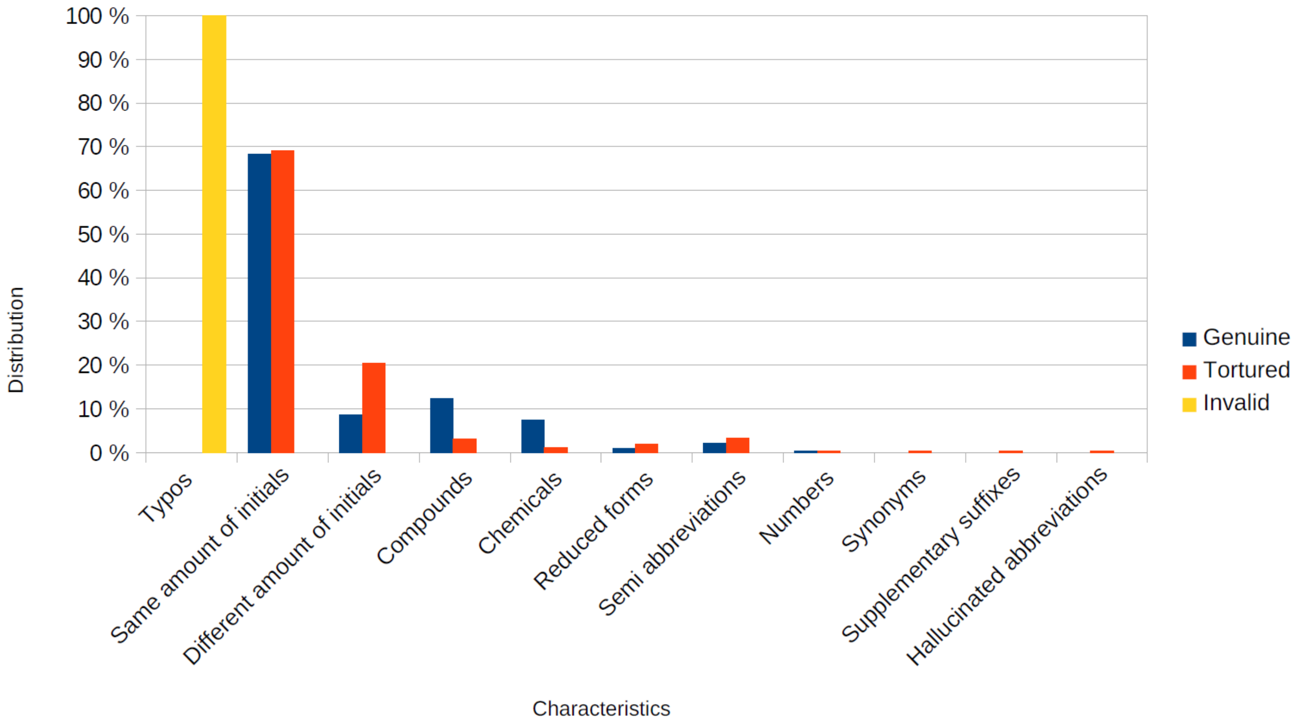
Annex 1: ditribution of publication years among the dataset



Annex 2: acronyms' characteristics categoristaion

Characteristics	Class		
	Genuine	**Tortured**	**Invalid**
Typos	-	-	*Tru_-Positive (TP)*
Same amount of initals	*Pain Dysfunction Syndrome (PDS)*	*Concealed Markov Display (HMM)*	-
Different amount of initials	*Centers for Diseases Control and Prevention (CDC)*	*Summed Up Direct Models (GLM)*	-
Compounds	*Intraocular Pressure (IOP)*	*Multidrug Safe (MDR)*	-
Chemicals	*3-Aminopropyltriethoxysilane (3-APTES)*	*Diethylenetriamine Pentaacetic ruinous (DTPA)*	-
Reduced forms	*Residual neural Networks (ResNets)*	*Lingering brain Organization (ResNet)*	-
Semi-abbreviations	*Gabor-HOG (GHOG)*	*Non-straight ARMA models (NARMA)*	-
Numbers	*Three-Dimensional (3D)*	*Fifth-Age (5G)*	-
Synonyms	-	*Human PC Interface/Cooperation*	-
Supplementary suffixes	-	*PC Assisted Determination (CADx)*	-
Hallucinated abbreviations	-	*Bolster Vector Machine (BVM)*	-

Annex 3: ditribution of acronyms' characteristics among the dataset



Annex 4: full results from the benchmark

Task	Recall	Precision	F1	MCC
Extraction	0.99	0.89	0.94	-
Classification	0.95	0.79	0.86	0.78
Tortured acronyms extraction	0.93	0.62	0.74	-