Imperial College London

Natural Language Processing based toolbox for detecting novelty in scientific research papers

Author: Yousef Nami

Supervisor: Dr Loïc Salles

1. Introduction

We live in a time where information is being produced at an increasing rate (see Figure 1). Though this trend is indicative of more accessibility to research, it also means that the literature review process is becoming more laborious. Even after precise keyword searches for relevant papers, researchers are left with hundreds of publications that don't reflect new research trends.

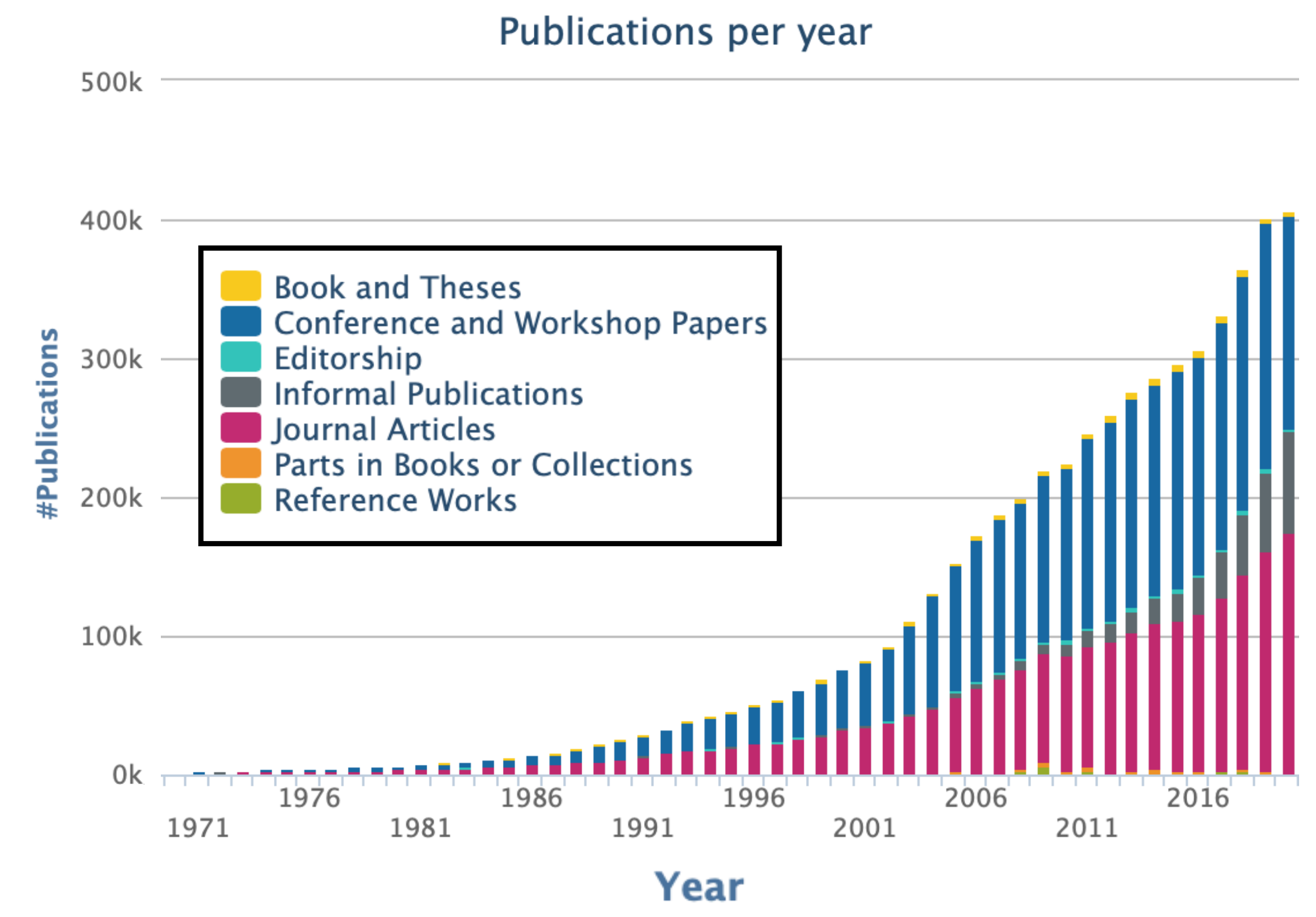


Figure 1: Major Computer Science publications since 70s [1]

This project aims to use deep semantic representations of text (made possible with the advent of deep learning) to determine the degree to which an unseen research paper is novel with respect to a pre-defined corpus in the field of Turbomachinery.

Two models (pseudo-supervised, unsupervised) are proposed to achieve this. This project is accompanied with a GitHub page [2] that includes Python-based Natural Language Processing (NLP) tools for extracting text from Open Access journals articles (from Elsevier, Core, IEEE and Hindawi) and PDFs, as well as pipelines for text processing, analysis and modelling.

2. Methodology

Text as vectors

GloVe word embeddings are used to represent text [3]. These are vector representations of text that capture semantic information, such as how close two words are in meaning (for example, $sausage$ is close to ham). They are also linear and additive (for example, $king - man + woman \approx queen$). Both properties are shown in Figure 2.

A statistical tool called TF-IDF. Each word is given a weighting that is rewarded if the word appears frequently in a document, but punished if for how many times it appears across other documents. This representation is based on word frequency, so interactions between words are not captured.

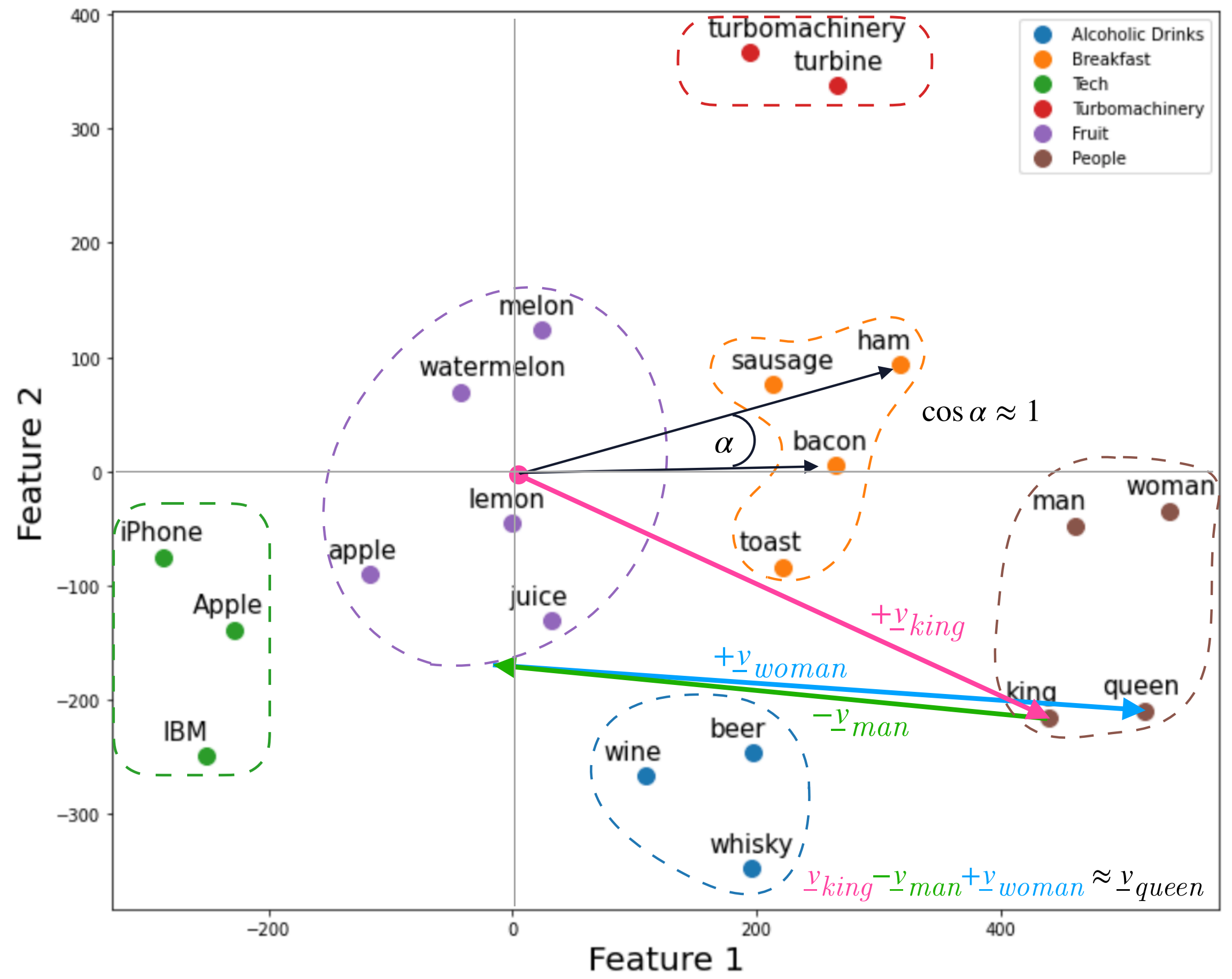
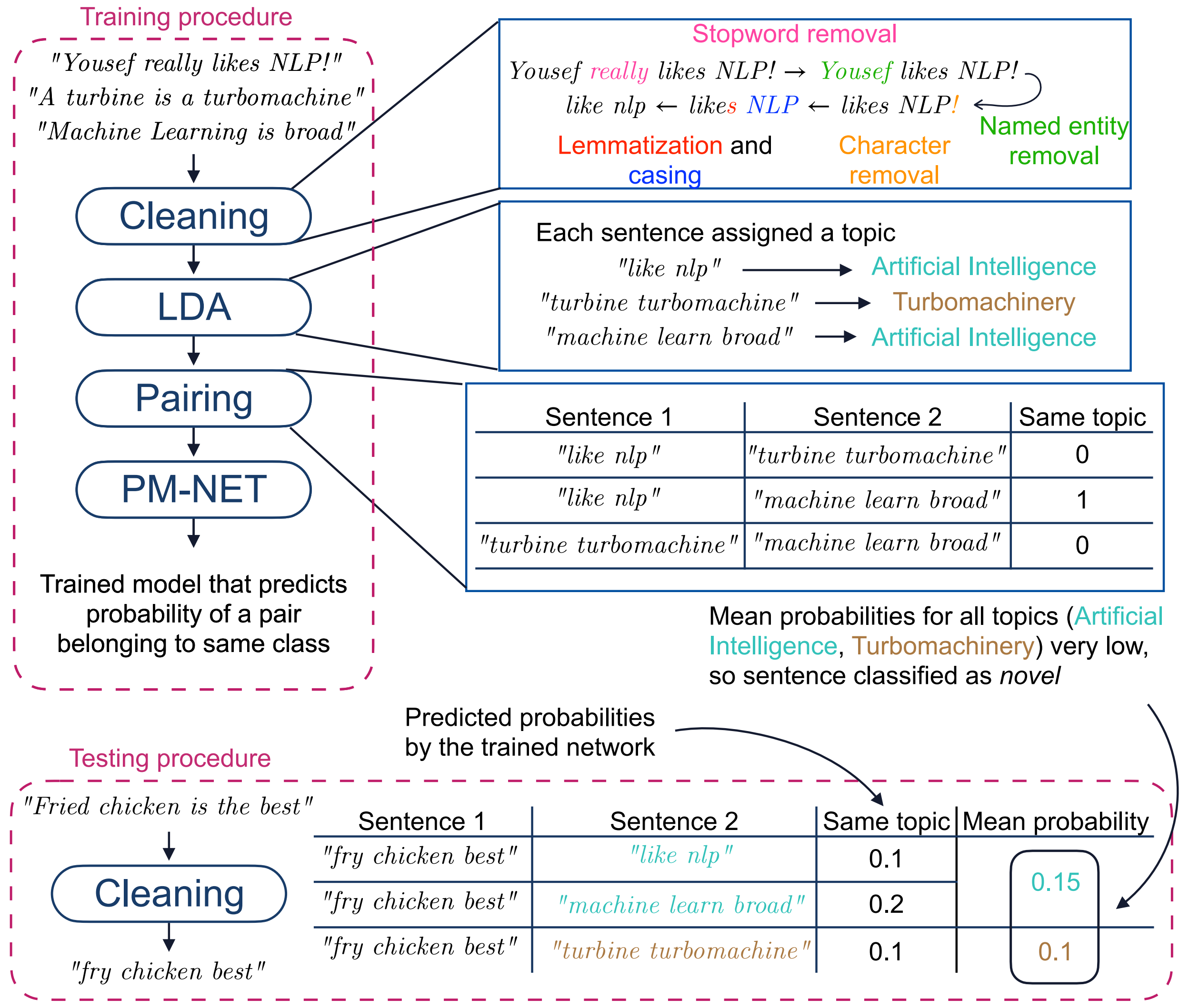


Figure 2: GloVe word embeddings plotted on a 2D plane showing additive behaviour of vectors and spatial closeness

Pseudo-supervised model (Model 1)

This model is based on the **Pairwise Matching Network (PM-NET 1)** by Qin et al. [4], which determines whether a sentence is novel or not. This idea is extended to document novelty. Each document is *sentencized,* then a pre-trained Latent Dirichlet Allocation (LDA) model predicts a topic for each sentence, creating a pseudo-supervised dataset. The percentage of novel sentences in a test document is its novelty score. A schematic of how this model behaves is shown in Figure 3 below.



Training procedure

"Yousef really likes NLP!"
"A turbine is a turbomachine"
"Machine Learning is broad"

Cleaning

LDA

Pairing

PM-NET

Trained model that predicts probability of a pair belonging to same class

Stopword removal

Yousef really likes NLP! → Yousef likes NLP!
like nlp ← likes NLP ← likes NLP!

Lemmatization and casing

Character removal

Named entity removal

Each sentence assigned a topic

"like nlp" → Artificial Intelligence
"turbine turbomachine" → Turbomachinery
"machine learn broad" → Artificial Intelligence

Sentence 1	Sentence 2	Same topic
"like nlp"	"turbine turbomachine"	0
"like nlp"	"machine learn broad"	1
"turbine turbomachine"	"machine learn broad"	0

Mean probabilities for all topics (Artificial Intelligence, Turbomachinery) very low, so sentence classified as *novel*

Predicted probabilities by the trained network

Testing procedure

"Fried chicken is the best"

Cleaning

"fry chicken best"

Sentence 1	Sentence 2	Same topic	Mean probability
"fry chicken best"	"like nlp"	0.1	0.15
"fry chicken best"	"machine learn broad"	0.2	
"fry chicken best"	"turbine turbomachine"	0.1	0.1

Figure 3: a high level schematic of the model's workings

Unsupervised model (Model 2)

Each document is represented by average TF-IDF or GloVe vectors. Then **IsolationForest** is used to detect anomalies, which are classed as novel. The intuition behind this is:

- A novel document is one that is anomalous
- Anomalies are typically very few
- **IsolationForest** works with highly imbalanced datasets [5]

3. Experiments

Datasets

Table 1: Project datasets and extraction methods

	Details	Extraction Method
Train	**ASME Turbomachinery** 80% of 26030 PDFs	Remote (linux device) PDF to text using `pdfminer`
Validation	Remaining 20% of above	
Test	**Hindawi Turbomachinery** SV and IJRM journals (2008—2020) stored in XML format	Custom XML parser
	Core Turbomachinery and IEEE Deep Learning 25 JSON papers each	User query search using publisher API
	9 Select PDFs on NLP	Local PDF to text

Implementation details

Model 1 is computationally and memory intensive, so the data pairing process was batched. The number of LDA topics was optimised using Jaccard and Coherence. The model was set to train on the College Computing cluster with 4 RTX6000 GPUs.

The unsupervised model was used on TF-IDF and GloVe represented documents. Its hyperparameters were optimised using the test data as ground truths.

Results and discussion

Model 1 did not train for more than 1000 sentences (out of 14 million), due to a time limit set by the College cluster system. This sample size is not unrepresentative due to higher validation accuracies and uncharacteristic training curves. The confidence of the optimised LDA model predictions are shown in Figure 4.

The results for Model 2 are shown in Figure 5. We can see that the model is not consistent at detecting novelty, since many points that appear very far from the cluster at not labelled as novel. Most of the Hindawi papers were classified as novel, which is a surprising result. Expert opinion is needed to determine the feasibility of this model beyond visual inspection.

Finally, the ASME corpus data quality is very low, with many incorrectly captured words. Most importantly, the GloVe vectors ignore many technical terms, such as $eulerian$, which are important for rich semantic representations of the documents.

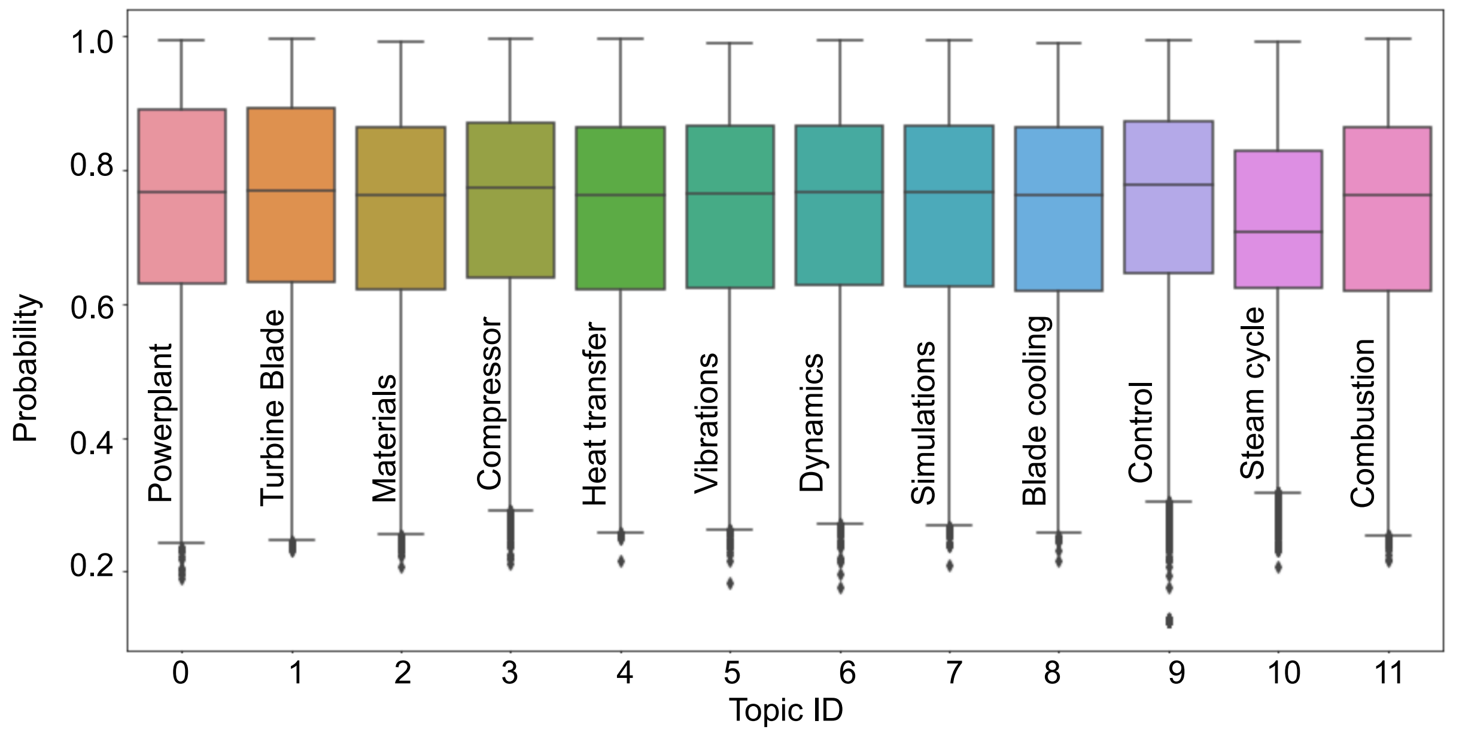


Figure 4: Box plot of LDA confidence in predicting sentence topics

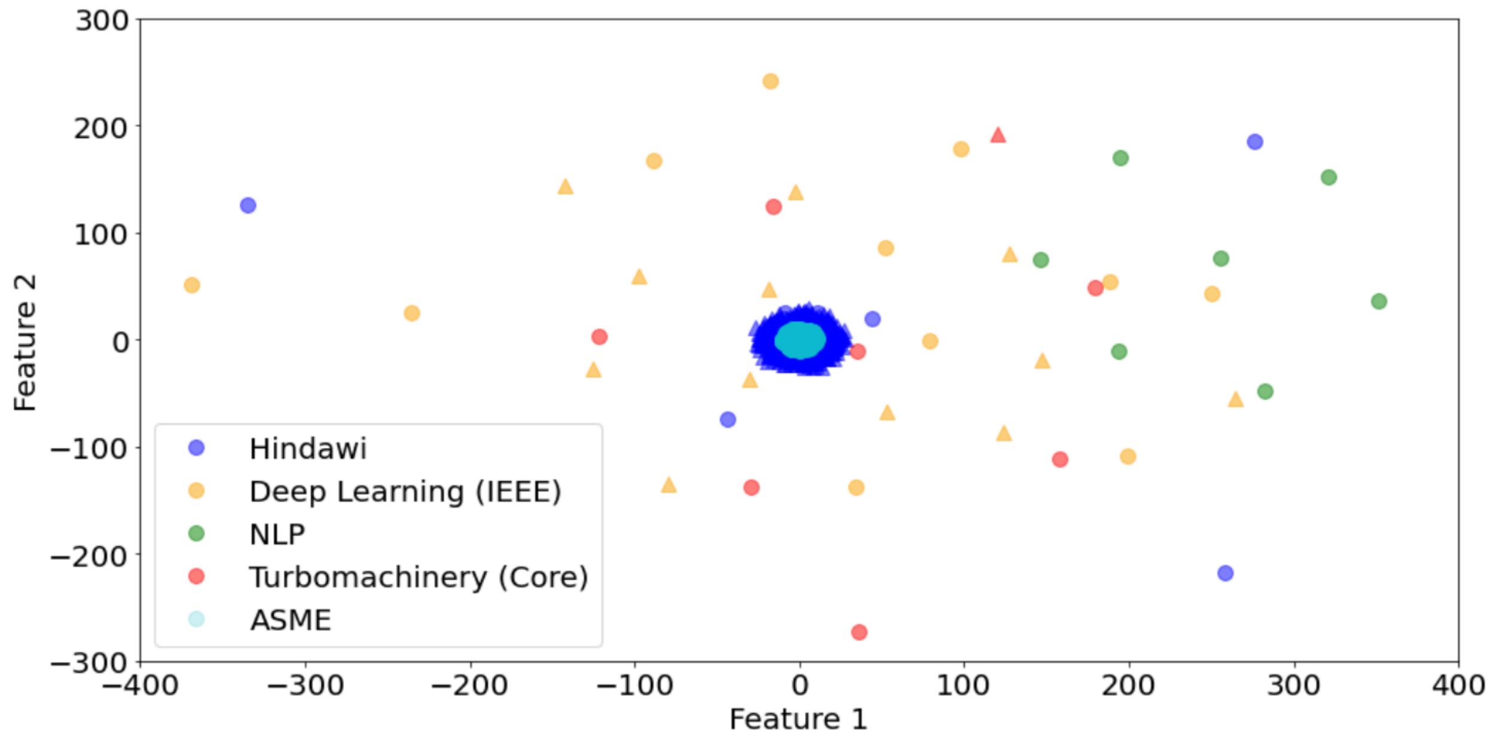


Figure 5: GloVe document representations with anomalies (triangle)

4. Conclusions

Model 1 is computationally and memory intensive, and thus could not train. It is best left for small-sized text documents. The use of LDA to label sentences with topics needs to be rigorously checked to ensure the validity of the method. The unsupervised model performed predictably, but was unable to consistently detect novel papers. TF-IDF is a good tool for filtering papers based on relevance before novelty detection.

Future developments should prioritise **Data Enhancement** (better text extraction, processing and correction), **Document Representation** (use of state-of-the-art document embedding techniques) and **redefining** novelty to leverage graphical structures in text.

References

[1]: dblp. Publications per year. Available from: dblp.org/statistics/publicationsperyear.html [Accessed 5th June 2021]

[2]: lsalles23. *ContentMining*. Available from: github.com/lsalles23/ContentMining

[3]: PJ, SR, MCD. *GloVe for Word Representation*. Available from: nlp.stanford.edu/projects/glove/ [Accessed 5th June 2021]

[4]: QQ, HW, LB. Text Class. with Novelty Detection. Available from: arxiv.org/abs/2009.11119 [Accessed 5th June 2021]

[5]: NY. *Anomaly Detec. in Sig. Movements*. Available from: github.com/namiyousef/Kin-Keepers/ [Accessed 5th June 2021]