Case example: Using NLP in research

PatentSBERTa

A Deep NLP based Hybrid Model for Patent Distance and Classification using Augmented SBERT

Hamid Bekamiri, Daniel Hain, Roman Jurowetzki
Aalborg University Business School
AI:Growth Lab, IKE

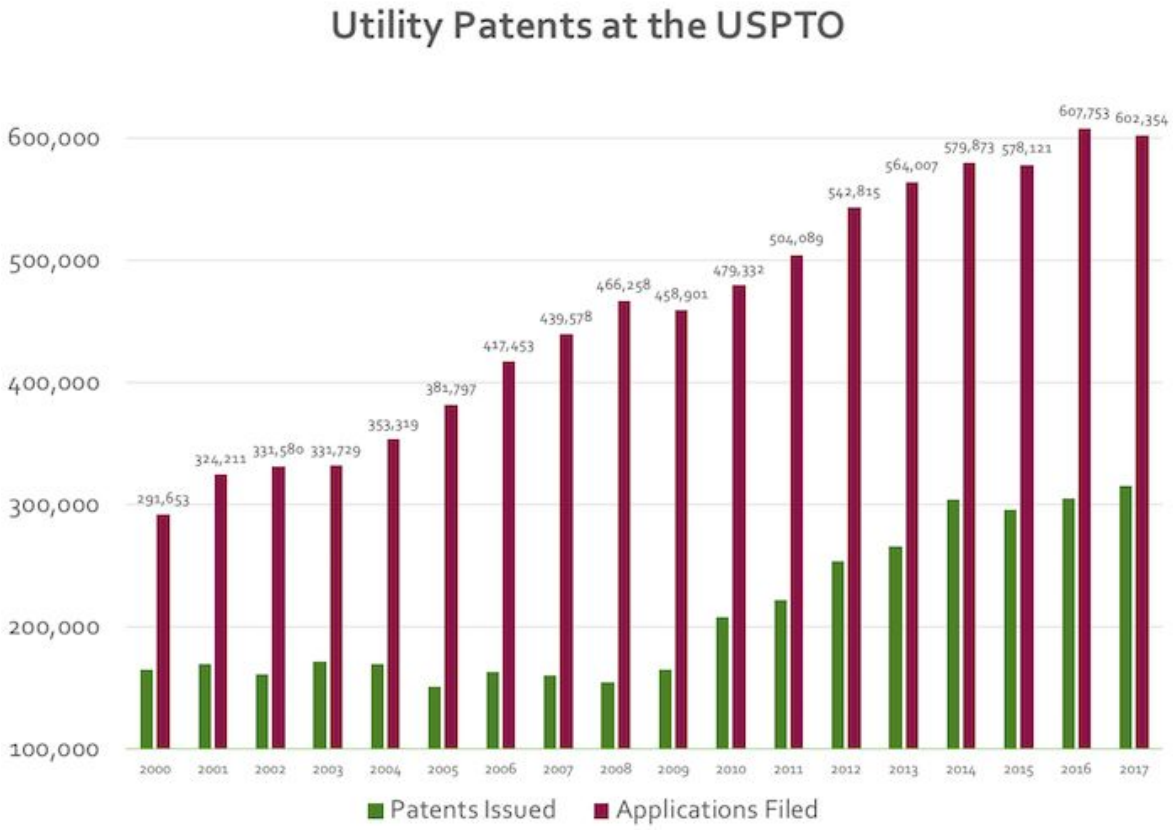


AI GROWTH LAB

BUSINESS SCHOOL
AALBORG UNIVERSITY

IKE
Innovation, Knowledge and Economic dynamics

Why patent semantic search and classification is important?



WHAT PATENTS COST

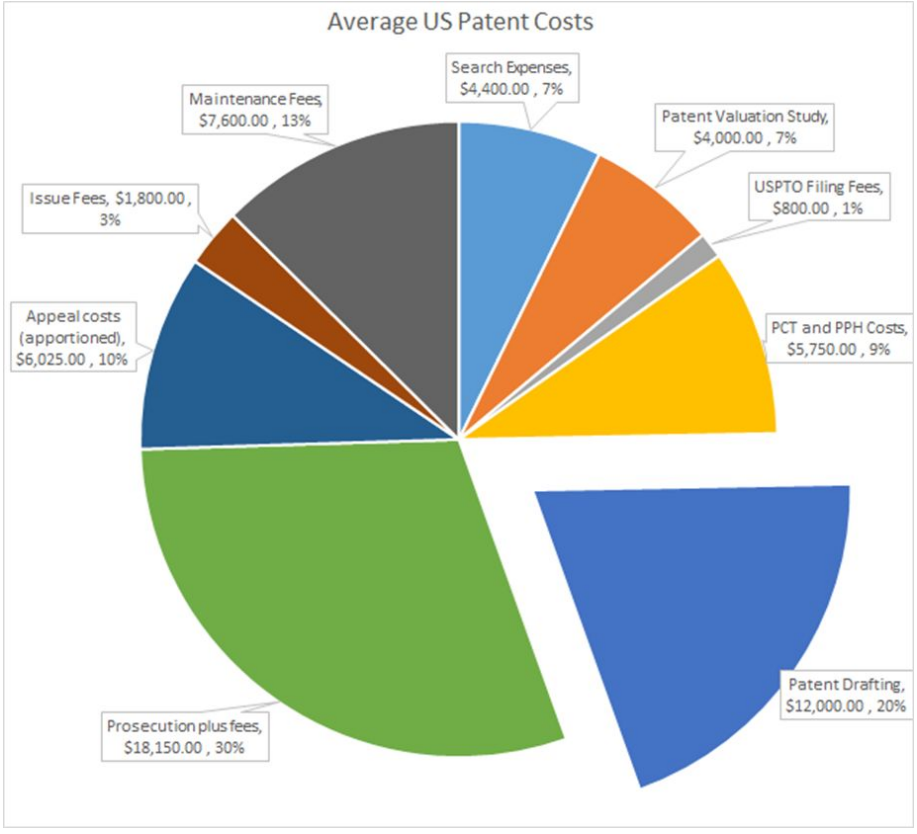
US NATIONAL AVERAGE

Notes:

1. All USPTO fees are based on Small Entity fees. Large entity fees are two times higher. These vary slightly from year to year.
2. All attorney's fees are average fees based on American Intellectual Property Lawyers Association bi-annual survey of 2019.
3. The total cost of $56,525 is an average cost of a patent in the US with a PCT filing. It reflects the average 4.2 office actions, a 75% probability of a Pre-Appeal Conference, and a 25% probability of a Full Appeal.

TOTAL COST: $56,525

Fees paid to USPTO	$12,100
Fees paid to Patent Attorney	$44,425



Patent search cost was more than 2.6 B$ just for 2019 in US Patent.

Patent litigation destroys over $60 billion in firm wealth each year.

Motivation for This Study
USPTO Strategic Plan

AI as a strategic focus

- 2018-2022 USPTO Strategic Plan:
  - Optimize development and delivery of information technology tools, including artificial intelligence and machine learning, for internal users of patent systems to ensure that they have the tools they need for a thorough search and examination.*

* Goal I: Optimize patent quality and timeliness; Objective 3: Foster innovation through business effectiveness

Challenges with AI

- AI is trained, not pre-programmed
- Performance of AI depends on quality data
- Models may not be generalizable
- Perception of a "black box"
- Expense of intellectual validation
- Models may require continuous updates

Motivation for This Study
USPTO Operational Goal

AI priorities for patents

Operational goal: leverage AI to improve effectiveness of examiners and the agency

- **AI for enhanced search**
  - Assessing prototype AI-based search enhancements
  - Investigating image recognition
- **CPC auto-classification**
  - Full CPC classification
  - Identification of CPC symbols associated with claims

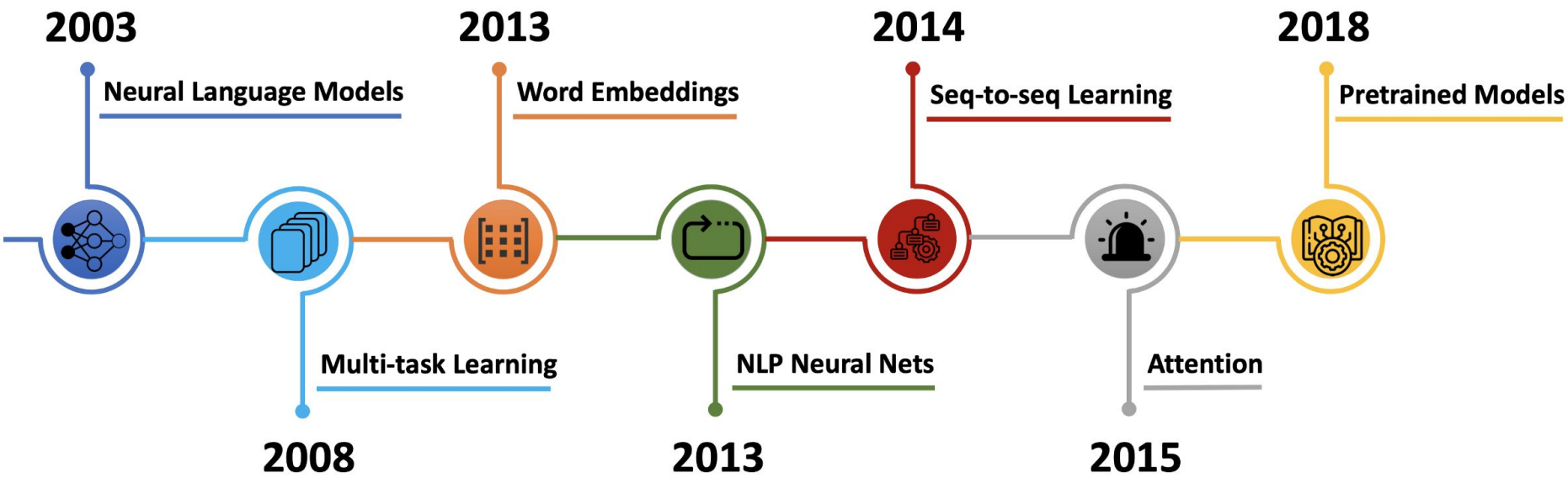
Strategy for reliable AI

- Curation of high-quality data is critical
- Apply solutions for validation and refinement
- Expand practical knowledge in AI
- Extensive outreach and market research
- AI is for augmentation
- Explainable AI

Motivation for This Study

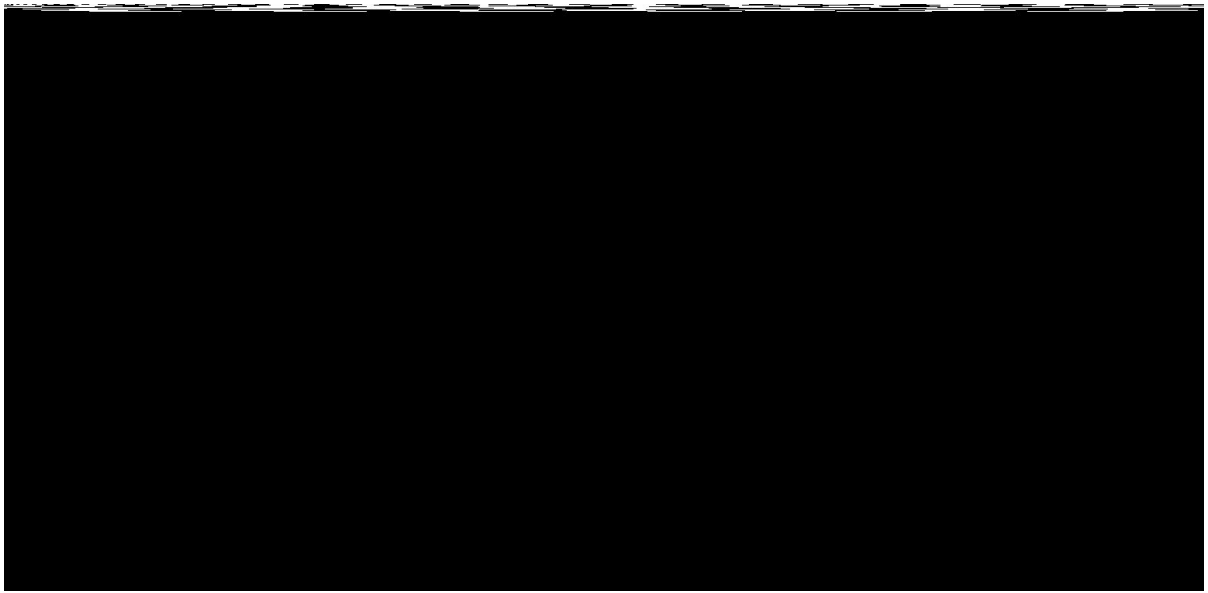
- Patent semantic search is a fundamental application in patent analytics.
- This task is a significant challenge for researchers because it is a multi-label classification task.
- With the increase in the number of patents, the design and implementation of automated methods to review these patents have increased in recent years (Abbas, et al., 2014).

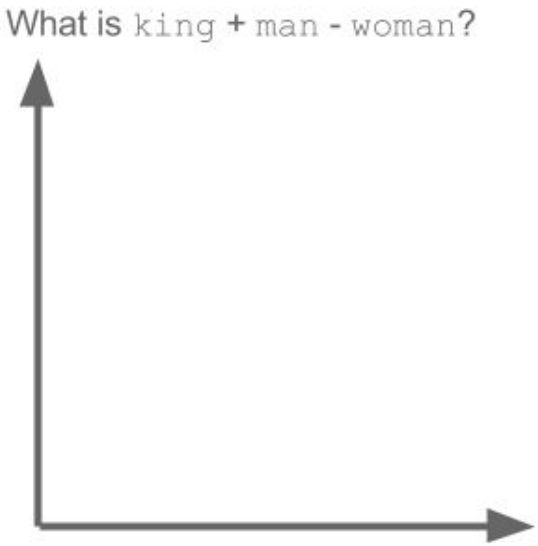
A Brief History of Natural Language Processing



https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-1-ffbcb937ebce

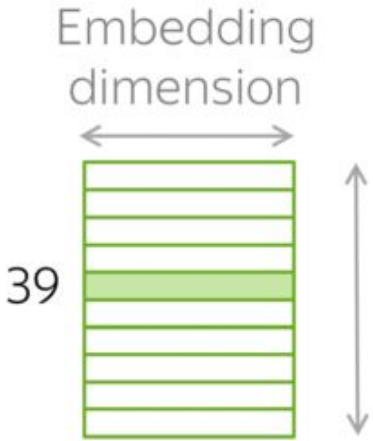
Word Embedding

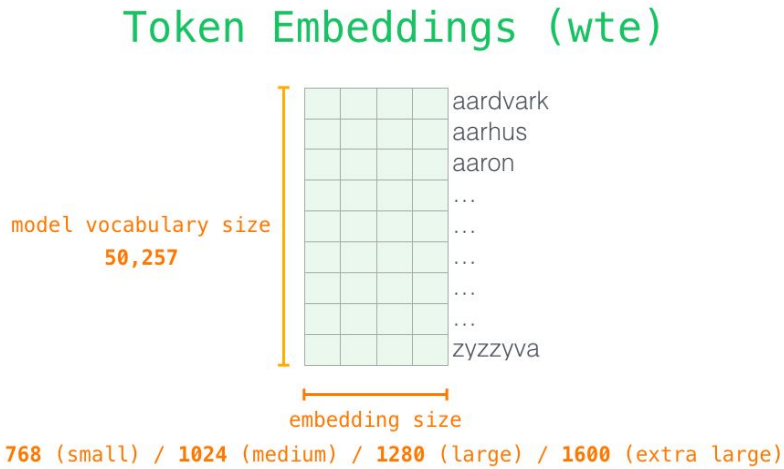




https://towardsdatascience.com/text-classification-with-nlp-tf-idf-vs-word2vec-vs-bert-41ff868d1794

Word Embedding





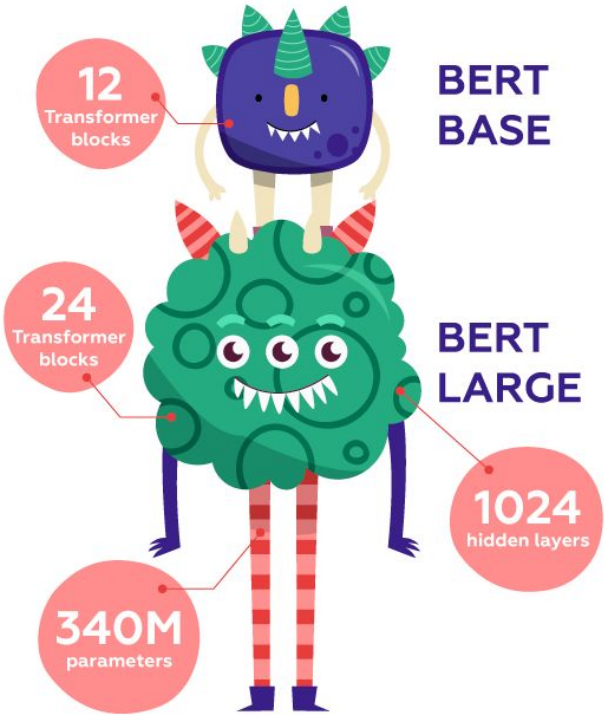
http://jalammar.github.io/illustrated-word2vec/


BERT

sciforce

BERT model at a glance

BERT comes in two sizes: BERT BASE, comparable to the OpenAI Transformer and BERT LARGE – the model which is responsible for all the striking results.



BERT is pre-trained on 40 epochs over:

3.3B word corpus

- **800M words** — BooksCorpus
- **2,5B words** — English Wikipedia

16 TPUs for training runs on

INPUT

BERT takes a sequence of words which keep flowing up the stack. Each layer applies self-attention, and passes its results through a feed-forward network, and then hands it off to the next encoder.

OUTPUT

The output of each position is a vector of size called hidden_size (768 in BERT Base). This vector can be used as the input for a classifier you choose.

80,5%

7,7% absolute improvement

The fine-tuned model for different datasets improves the GLUE benchmark to 80.5%

86,7%

4,6% absolute improvement

MultiNLI accuracy to 86.7%

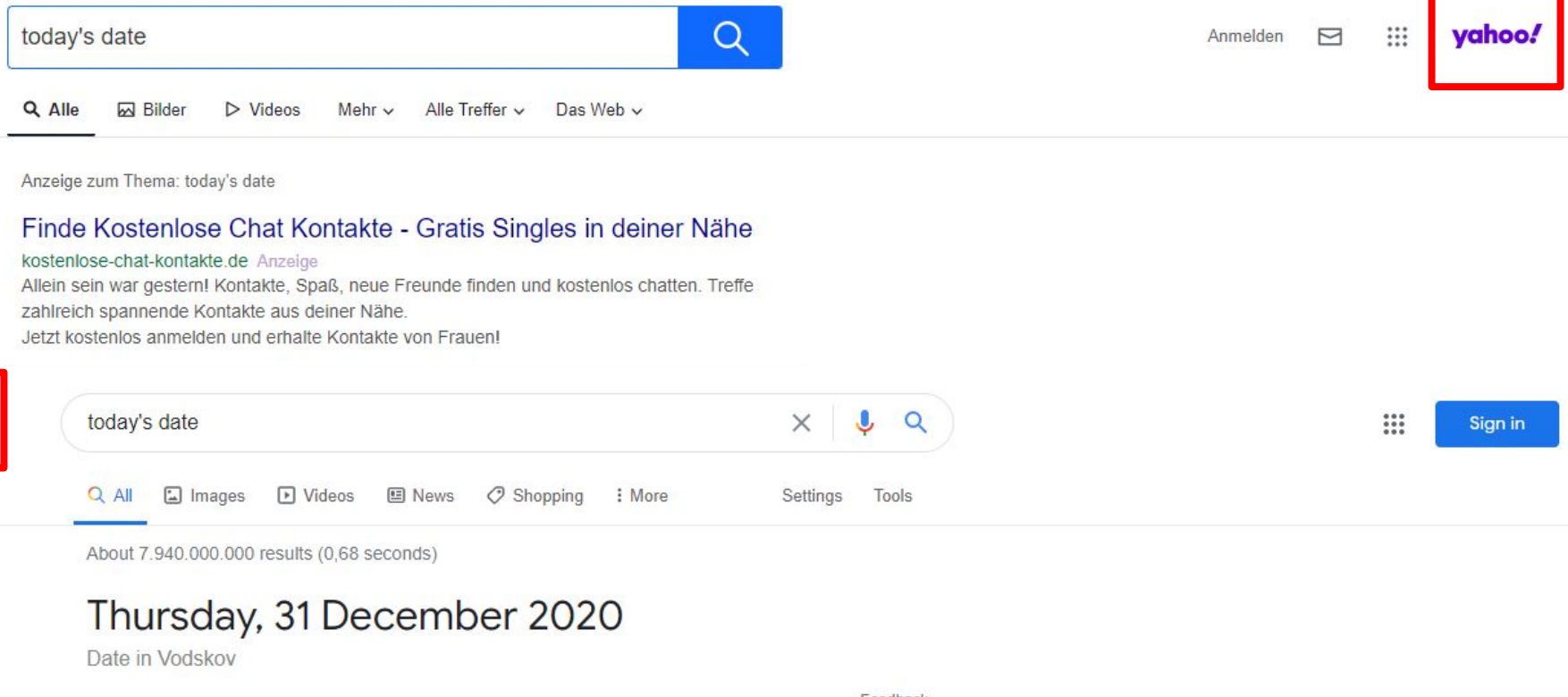
93,2%

1,5% absolute improvement

the SQuAD v1.1 question answering Test F1 to 93.2

So on over a total of 11 language tasks

Difference between Google (BERT) and Yahoo search results



Industry and Transformer Models

Use-case	Input example	Output
Text classification monzo	"I'm very unhappy about my new credit card"	Priority: **high** Sentiment: **unhappy** Department: **credit cards**
Information extraction Chegg	"Georges Washington became the first president in 1781"	Washington = **person** 1781 = **date** topic = **history**
Question-answering Bing	"What is the color of Kim Kardashian's hair?"	Kim Kardashian's hair is **blue**
Text generation / summarization	APPLE +10 points	Apple's share grew by 10 points today in what marks the highest rise since…
Conversational Square	When will my package arrive?	It will arrive on Oct 25th

https://www.youtube.com/watch?v=ecEwekuJ2

Parts of a Patent

A patent application is issued as a patent after it is allowed by the US Patent and Trademark Office (USPTO). All claims are novel and non-obvious.

Patent Number

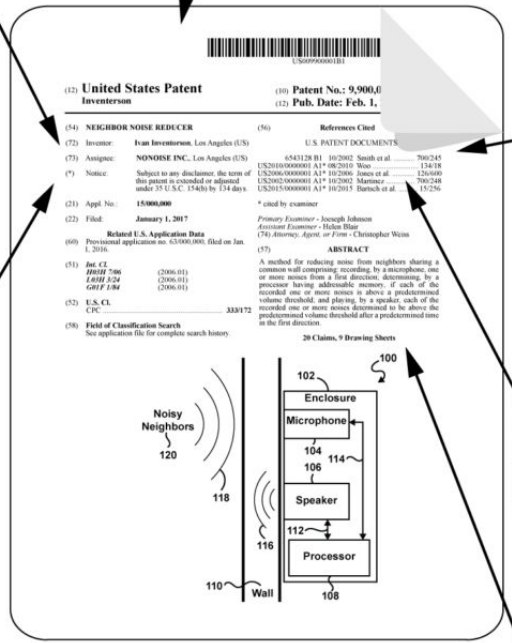
Patents have a publication number in the format US X,XXX,XXX. An issued patent allows you to exclude others from making, using, or selling your invention in the United States based on your claims.

Owners

The inventor owns the rights to a patent unless it is assigned to another person or company.

Time

A US patent lasts for twenty years from the date of filing. If a patent takes a long time to be grated, due to delays, the USPTO can add extra time to the length of the patent.



(12) **United States Patent**
Inventerson

(10) **Patent No.: 9,900,0**
(12) **Pub. Date: Feb. 1,**

(54) **NEIGHBOR NOISE REDUCER**

(72) Inventor: **Ivan Inventorson**, Los Angeles (US)

(73) Assignee: **NONOISE INC.**, Los Angeles (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 134 days.

(21) Appl. No.: **15/000,000**

(22) Filed: **January 1, 2017**

Related U.S. Application Data

(60) Provisional application no. 63/000,000, filed on Jan 1, 2016.

(51) *Int. Cl.*
H04R 7/06 (2006.01)
L01H 3/24 (2006.01)
G01F 1/84 (2006.01)

(52) **U.S. Cl.**
CPC **333/172**

(58) **Field of Classification Search**
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6543128 B1 10/2002 Smith et al. 700/245
US2010/0000001 A1* 08/2010 Woo 134/18
US2006/0000001 A1* 10/2006 Jones et al. 126/600
US2002/0000001 A1* 10/2002 Martinez 700/248
US2015/0000001 A1* 10/2015 Bartsch et al. 15/256

* cited by examiner

Primary Examiner - Joseph Johnson
Assistant Examiner - Helen Blair
(74) *Attorney, Agent, or Firm* - Christopher Weiss

(57) **ABSTRACT**

A method for reducing noise from neighbors sharing a common wall comprising recording, by a microphone, one or more noises from a first direction; determining, by a processor having addressable memory, if each of the recorded one or more noises is above a predetermined volume threshold; and playing, by a speaker, each of the recorded one or more noises determined to be above the predetermined volume threshold after a predetermined time in the first direction.

20 Claims, 9 Drawing Sheets

References

A listing of relevant references cited by the inventor and found by the Examiner during a search.

Examiners

The Examiners that reviewed the patent application are listed along with any patent attorneys hired by the inventors.

Claims

The total number of claims and sheets of drawings are listed in this section. The claims in a patent define the scope of protection.

US 20190080260A1

(19) **United States**

(12) **Patent Application Publication**
Acuna Agost et al.

(10) **Pub. No.: US 2019/0080260 A1**
(43) **Pub. Date: Mar. 14, 2019**

(54) **MACHINE LEARNING METHODS AND SYSTEMS FOR PREDICTING ONLINE USER INTERACTIONS**

(71) Applicant: **Amadeus S.A.S.**, Biot (FR)

(72) Inventors: **Rodrigo Acuna Agost**, Golfe Juan (FR); **Alejandro Ricardo Mottini D'Oliveira**, Nice (FR); **David Renaudie**, Valbonne (FR)

(21) Appl. No.: **15/704,320**

(22) Filed: **Sep. 14, 2017**

Publication Classification

(51) **Int. Cl.**
G06N 99/00 (2006.01)
G06N 5/02 (2006.01)
G06Q 30/02 (2006.01)

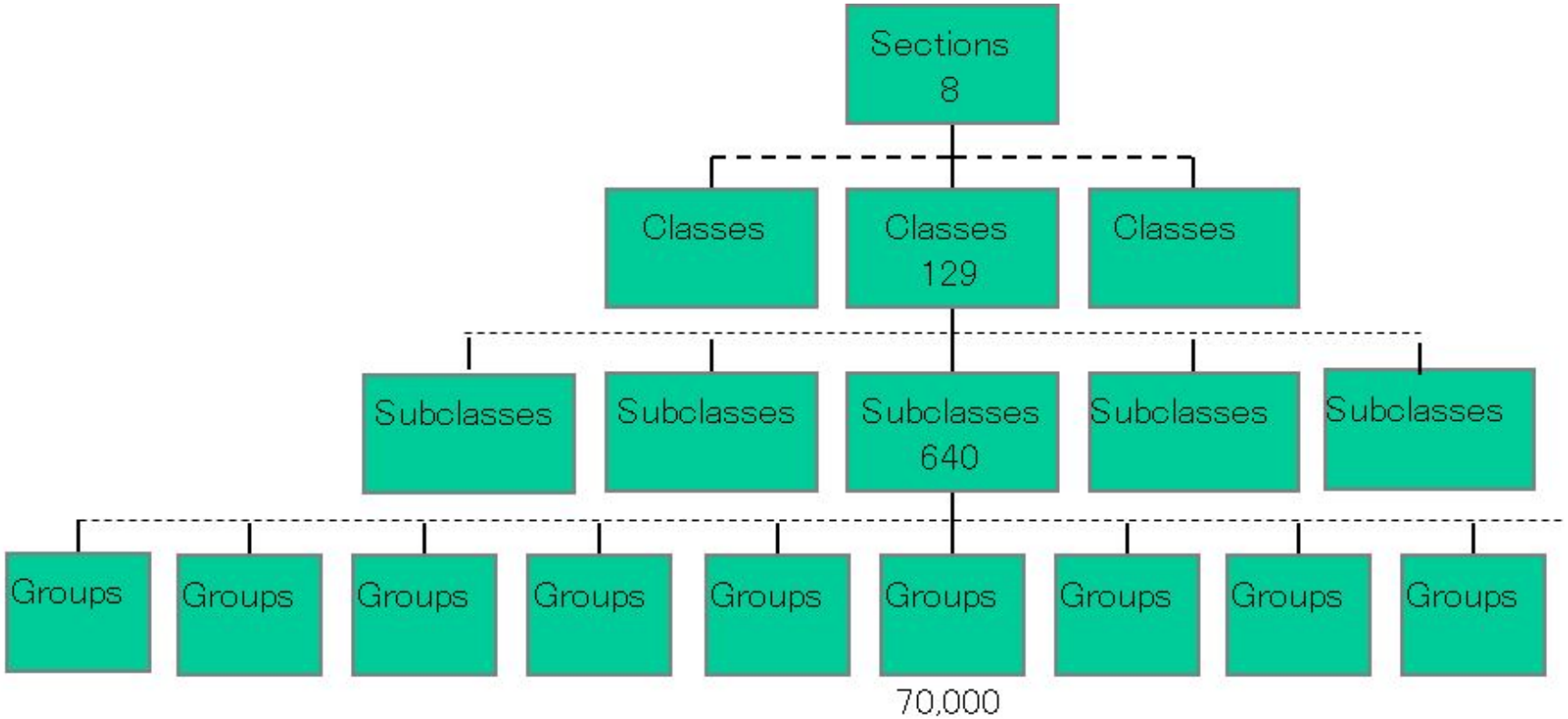
(52) **U.S. Cl.**
CPC ***G06N 99/005*** (2013.01); ***G06Q 30/0275*** (2013.01); ***G06Q 30/0242*** (2013.01); ***G06N 5/022*** (2013.01)

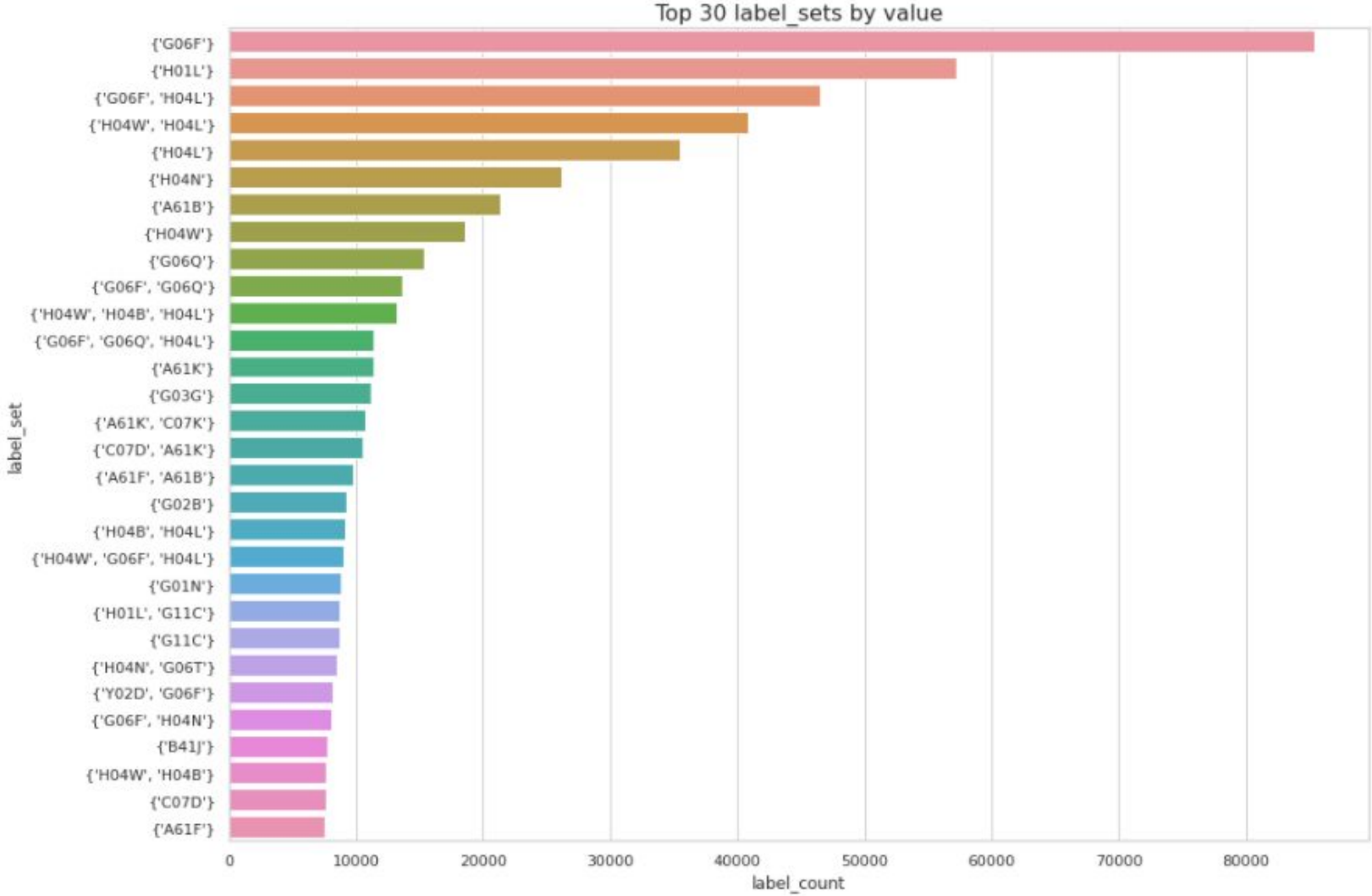
(57) **ABSTRACT**

Methods and computing apparatus for retrieving records relating to content placement events and records relating to user interaction events. A set of enriched training feature vectors is computed from raw feature values, and used with interaction event tags to train a machine learning model. Information is received relating to an online content placement slot and information is received relating to a user to whom content within the online content placement slot will be displayed. An enriched estimation feature vector is computed based upon a content item selected for placement within the online content placement slot, the information relating to the user, and the information relating to the online content placement slot. A machine learning model is executed to determine an estimate of likelihood of the user interacting with the selected content item, based upon the enriched estimation feature vector.

Numbers vs. Text

Hierarchical Structure of Patent Classification

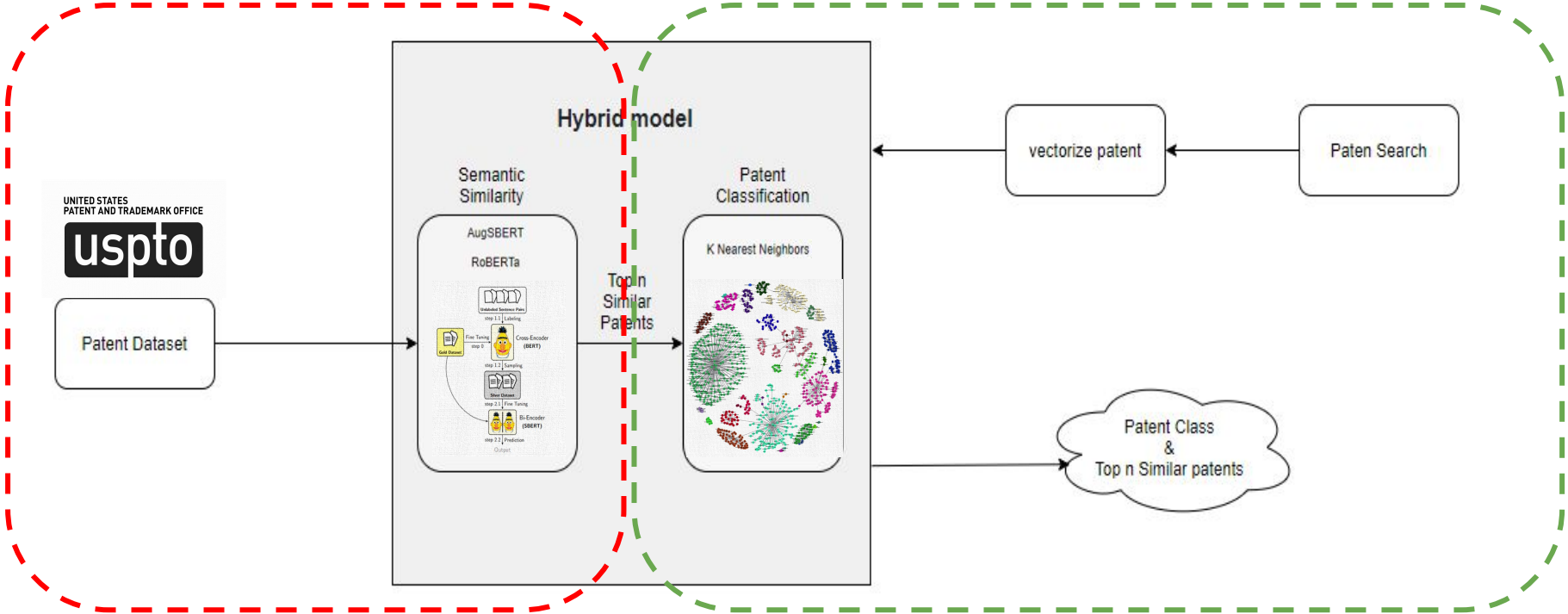



Top 30 label_sets by value
label_set
{'G06F'}
{'H01L'}
{'G06F', 'H04L'}
{'H04W', 'H04L'}
{'H04L'}
{'H04N'}
{'A61B'}
{'H04W'}
{'G06Q'}
{'G06F', 'G06Q'}
{'H04W', 'H04B', 'H04L'}
{'G06F', 'G06Q', 'H04L'}
{'A61K'}
{'G03G'}
{'A61K', 'C07K'}
{'C07D', 'A61K'}
{'A61F', 'A61B'}
{'G02B'}
{'H04B', 'H04L'}
{'H04W', 'G06F', 'H04L'}
{'G01N'}
{'H01L', 'G11C'}
{'G11C'}
{'H04N', 'G06T'}
{'Y02D', 'G06F'}
{'G06F', 'H04N'}
{'B41J'}
{'H04W', 'H04B'}
{'C07D'}
{'A61F'}
0
10000
20000
30000
40000
50000
60000
70000
80000
label_count

Text-based Patent Classification: Related Work

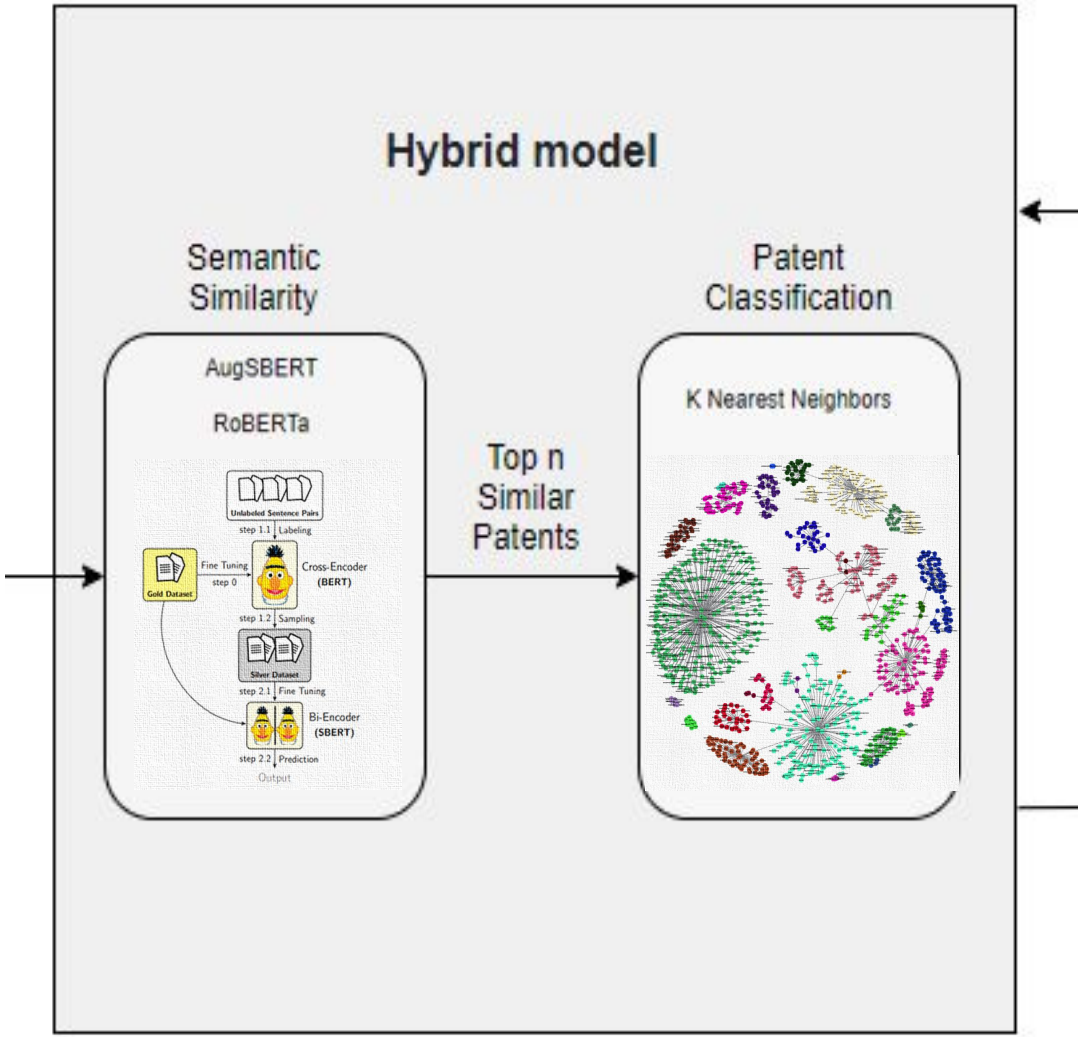
Method/Refrences	*Text Data*	*N Patents_Train*	*N Patents_Test*	*F1*	*Precision*	*Recall*	*EVAL*	*Number_Class*
ULMFiT SVM (Hepburn, 2018)	ALTA+WIPO	45,150	30,100	**78**	N/A	N/A	Score	8 labels
BiLSTM (Hu et al., 2018)	IPC+CLEF-IP	90,665	2,679	**64**	N/A	N/A	Top_1	96 labels
TF-ICF (Lim & Kwon, 2016)	Claim	N/A	564,793	N/A	**59**	N/A	Score	650 labels
	Titles, Abstracts	N/A	564,793	N/A	**88**	N/A	Score	
DeepPatent (Li et al., 2018)	IPC+Title+Abstract	580,546 + 161,551	1,350	N/A	**84**	N/A	Top_1	637 labels
	IPC+Title+Abstract	2,000,147	49,900	N/A	**74**	N/A	Top_1	
	IPC+Title+Abstract	580,546 + 161,551	1,350	**55**	**46**	**75**	Top_1	
	IPC+Title+Abstract	2,000,147	49,900	**< 43**	**< 35**	**< 74**	Top_5	
PatentBERT (Lee & Hasing, 2020)	IPC+Title+Abstract	1,950,247	49,670	**65**	**81**	**54**	Top_1	656 labels
	IPC+Title+Abstract	1,950,247	49,670	**45**	**30**	**86**	Top_5	
	CPC+Claim	1,950,247	49,670	**67**	**84**	**55**	Top_1	
	CPC+Claim	1,950,247	150,000	**67**	**84**	**55**	Top_1	
	CPC+Claim	1,950,247	150,000	**81**	N/A	N/A	Score	8 labels
(Hain et al., 2021)	Titles, Abstracts	1,000,000	10,000	**52**	**54**	**53**	Top_1	637 labels

PatentSBERTa workflow



Part 1:
Semantic Similarity

- Using Augmented SBERT for calculating **Sentence embedding**
- SBERT **increase the performance** and accuracy of BERT for **semantic search** in sentence level



Why SBERT?

- SBERT performs particularly well for **semantic similarity** tasks

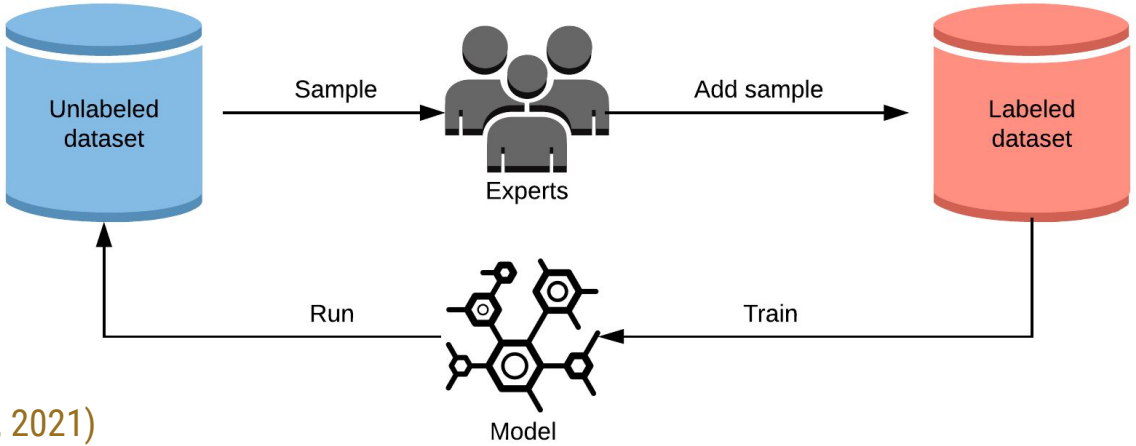
Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
InferSent - Glove	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18	74.92	**76.69**	71.22
SBERT-NLI-base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-NLI-large	72.27	**78.46**	**74.90**	80.99	76.25	**79.23**	73.75	76.55
SRoBERTa-NLI-base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-NLI-large	**74.53**	77.00	73.18	**81.85**	**76.82**	79.10	74.29	**76.68**

Table 1: Spearman rank correlation ρ between the cosine similarity of sentence representations and the gold labels for various Textual Similarity (STS) tasks. Performance is reported by convention as $\rho \times 100$. STS12-STS16: SemEval 2012-2016, STSb: STSbenchmark, SICK-R: SICK relatedness dataset.

Source: Reimers and Gurevych (2019)

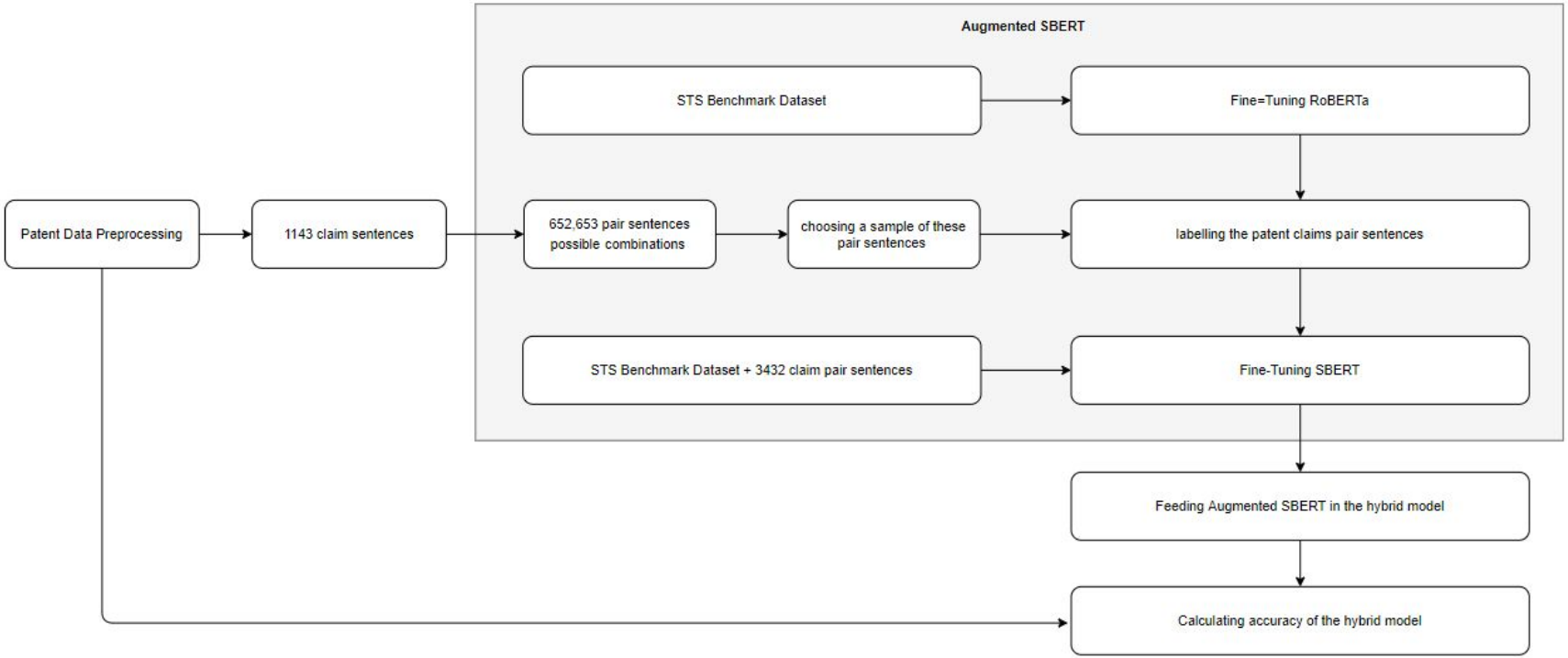
Why Augmented SBERT?

- BERT-based large pretrained models performance **in-domain** is still debatable (Pota et al., 2021).
- Experimental work indicates: Min. 1k–3k **In-domain labeled data** needed.
- Yet, labelling data is e**xpensive & time-consuming**
- **Self-supervis**



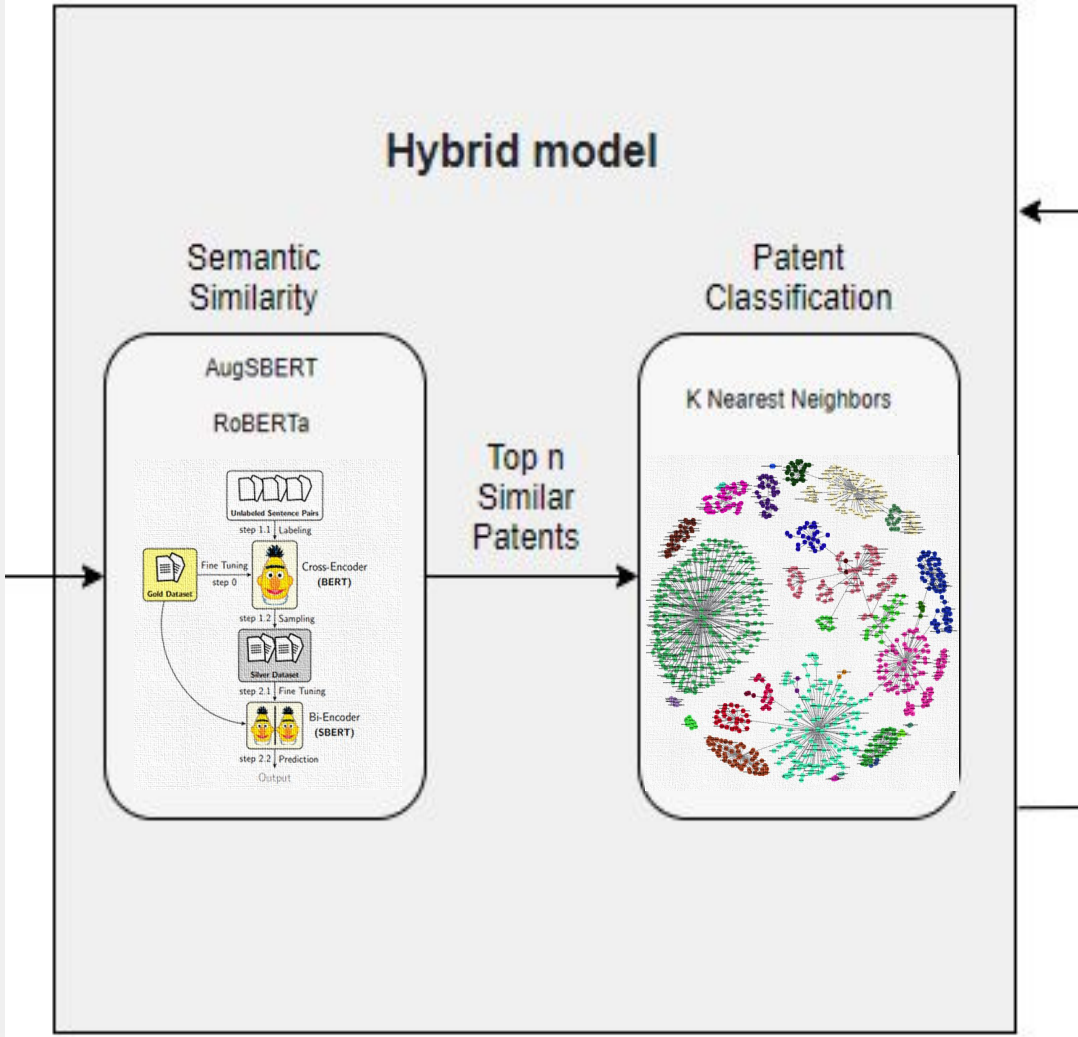
(Source: Thakur, N. et al., 2021)

Augmented SBERT workflow



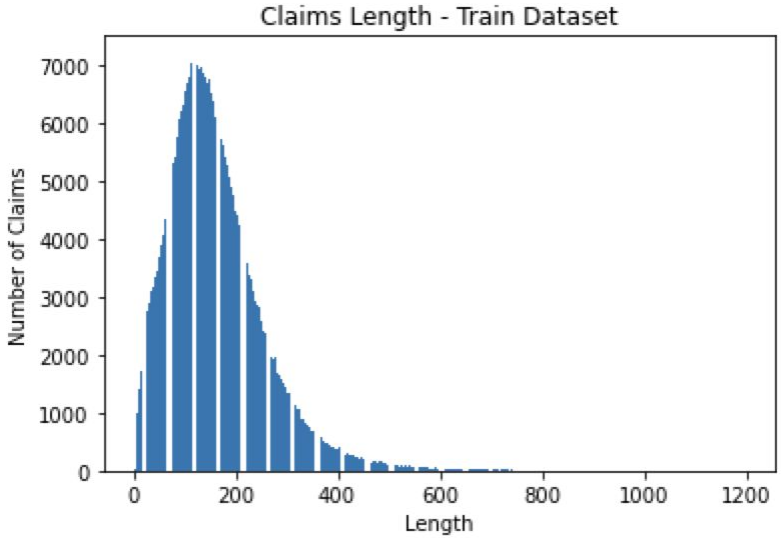
Part 2 Classification & Search

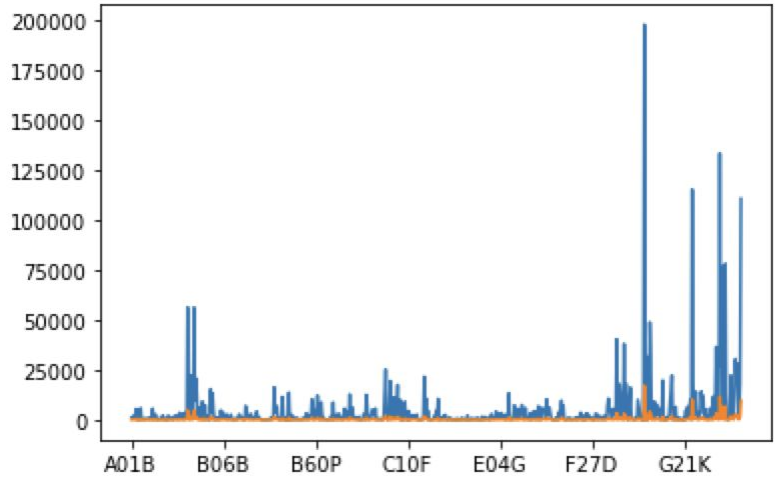
- We find N top **similar claims** and assign CPC classes based on the labels of n**earest claims** through a sigmoid function.
- The model calculate classification metrics for **different values** of K.



Data

- Datasource: Google Patent
- Timeframe: 2013-2017
- N patent: 1,492,294
- N Test Dataset: 119,384





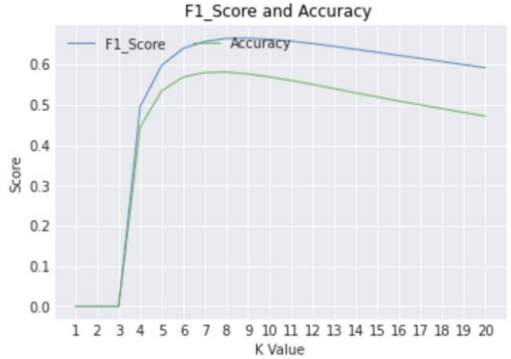
Example: KNN results

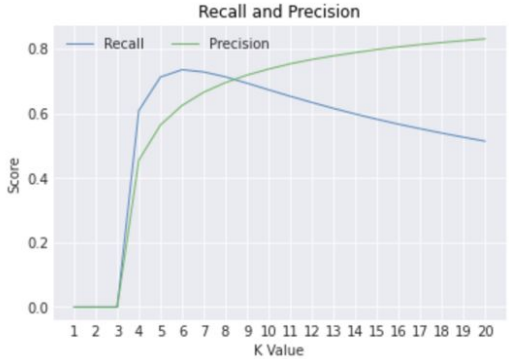
New_claim: An apparatus, comprising: an array of non-volatile memory cells including a plurality of sections each with a plurality of rows; and a controller configured to: move data stored in a first portion of the array in a first particular order based on a first wear leveling algorithm, wherein the first particular order includes moving data from a first row of a first section of the array to a second row of the first section of the array via a first sensing component stripe, wherein the first sensing component stripe is only coupled to rows of the first section of the array; and move data stored in a second portion of the array in a second particular order different than the first particular order based on a second wear leveling algorithm different from the first wear leveling algorithm and move data from a second section of the array to the first row of the first section of the array, via the first sensing component stripe and a second sensing component stripe, to create an open row in the second section of the array for wear leveling.

	top_claim_ids	cosine_similarity	claims	patent_id	text	section_id	subsection_id	group_id	subgroup_id
0	10606970	0.7243	A computer-implemented method for statistical ...	10606736	A computer-implemented method for creation of ...	G	G06	G06F	G06F11/3612
1	10606874	0.6766	A method implemented in a computer infrastruct...	10606738	A method, comprising: receiving results from a...	H	H04	H04L	H04L2209/56
2	10606858	0.6512	A method for presenting content based on a gen...	10606739	A device, comprising: a memory; and one or mor...	G	G06	G06F	G06F11/3684
3	10606995	0.6468	A character input device, comprising: an opera...	10606740	A system comprising: a processor comprising a ...	G,G,G	G06,G06,G06	G06F,G06F,G06F	G06F12/0207,G06F2212/455,G06F9/4881
4	10607072	0.6392	A method for computerized authentication of a ...	10606741	A process performed by a computing device for ...	G	G06	G06F	G06F3/0673

Results & Metrics

Method/Refrences	*Text Data*	*N Patents_Train*	*N Patents_Test*	*F1*	*Precision*	*Recall*	*EVAL*	*Number_Class*
	IPC+Title+Abstract	1,950,247	49,670	**65**	**81**	**54**	Top_1	656 labels
	IPC+Title+Abstract	1,950,247	49,670	**45**	**30**	**86**	Top_5	
	CPC+Claim	1,950,247	49,670	**67**	**84**	**55**	Top_1	
	CPC+Claim	1,950,247	150,000	**67**	**84**	**55**	Top_1	
PatentBERT (Lee & Hasing, 2020)	CPC+Claim	1,950,247	150,000	**81**	N/A	N/A	Score	8 labels
	CPC+Claim	1,492,294	119,384	**66.48**	**74**	**60**	Score	663 labels
PatentSBERTa	CPC+Claim	1,492,294	119,384	**82.44**	**79**	**90**	Score	8 labels





Q&A, Good Bye, & Follow up

Check out & use our stuff:

- arXiv Paper: https://arxiv.org/abs/2103.11933
- **Github:** https://github.com/AI-Growth-Lab/Patent-Classification
- **Hugginface:** https://huggingface.co/AI-Growth-Lab/PatentSBERTa → Demo & Tutorial
- arXiv of prior paper: https://arxiv.org/abs/2003.12303
- Application paper: https://academic.oup.com/icc/article-abstract/29/5/1233/5923785
- Application demo: http://localhost:8501/

Reach out:

Daniel Hain (dsh@business.aau.dk) @Daniel_S_Hain daniel-hain
Associate Professor (Data Science & Innovation Economics), AI:Growth Lab, AAUBS

Roman Jurowetzki (roman@business.aau.dk)
Associate Professor (Data Science & Innovation Economics), AI:Growth Lab, AAUBS

Hamid Bekamiri (hamidb@business.aau.dk)
PostDoc (ML & Deep NLP), AI:Growth Lab, AAUBS

Thank you!

UNITED STATES
PATENT AND TRADEMARK OFFICE

CLAAUDIA

In this study, we used the patent view dataset. For exploring data we utilized SQLite on the AI AAU Cloud. The research dataset was about 4M patent claims. We downloaded the patents dataset from PatentsView for 2010-2020.

AI AAU Cloud has 3 nodes which each compute node is an NVIDIA DGX2 machine that contains 16 Tesla Volta V100 GPUs (32GB RAM each). Overall, AI AAU Cloud has 1.5TB of RAM and 30TB of NVME (SSD) for scratch space. Total GPU computation power is around 6 Petaflops.

AI Growth Lab

1. Summarization demo
2. Web Scraping demo
3. Patent search demo

https://blogs.nvidia.com/blog/2021/11/22/gaugan2-ai-art-demo/?ncid=so-link-483570-vt16#cid=nr01_so-link_en-us

Image captioning

Providing a natural language description of the content within an image