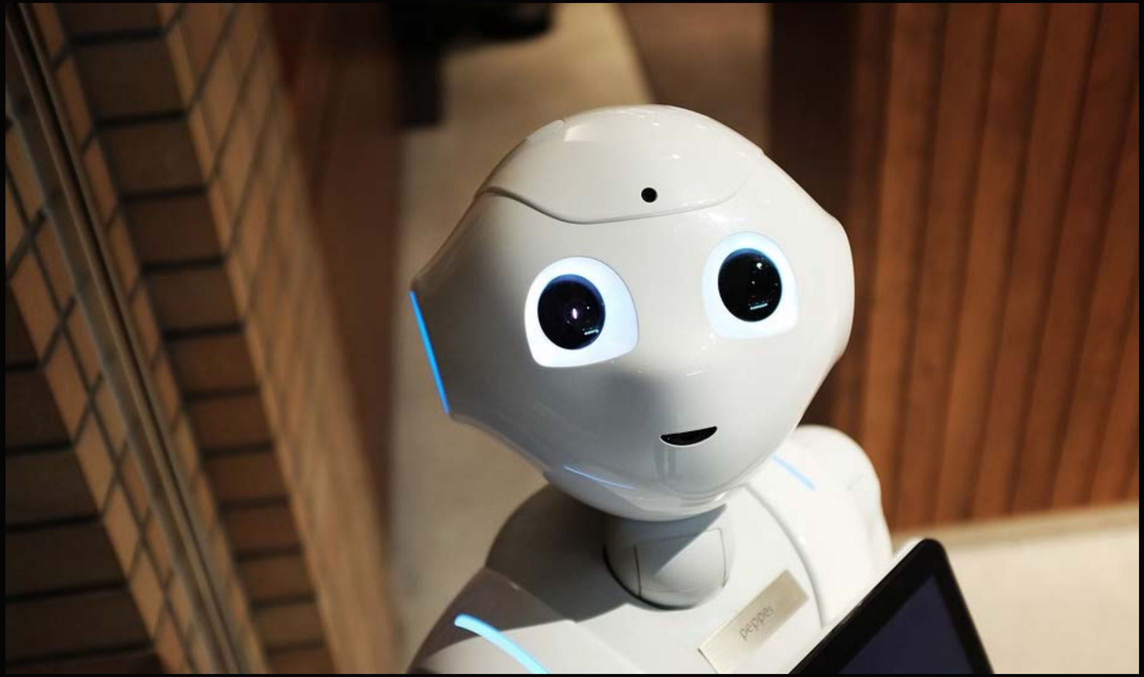
ROBERT GORDON UNIVERSITY ABERDEEN

Can machines read sentences like Geoscientists do?



Paul H. Cleverley Ph.D.

Associate Lecturer

Robert Gordon University, Aberdeen, Scotland UK.

FORCE – Machine Learning on Subsurface Data Symposium 20th Sep 2018

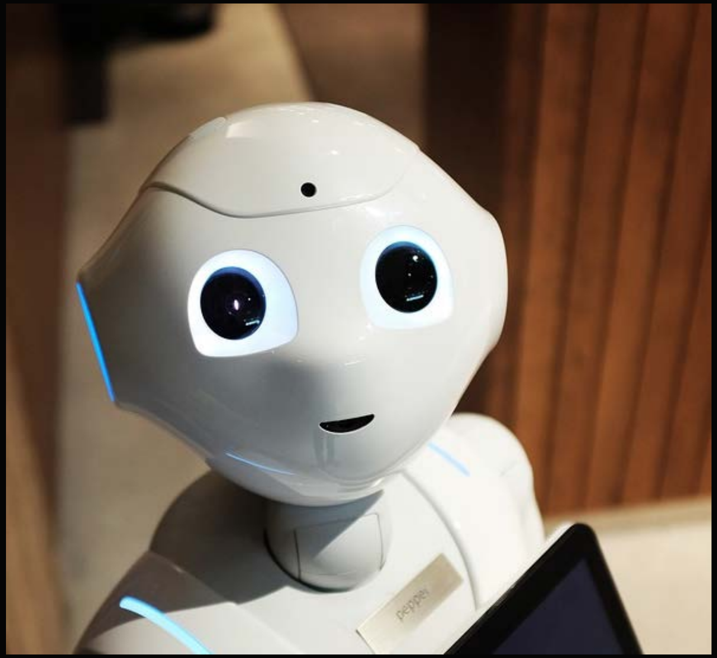
Can machines read sentences like geoscientists do?

Assumptions – The 3 laws[1] of geoscience text analytics

1. **Law of Non-Equivalence**: Algorithms will not read or understand sentences in the same way as *geoscientists*, because they are not geoscientists. It will be different, for better and for worse. Algorithms will have their own biases – but they may differ from our own ones which may assist and be helpful.

2. **Law of Human Cognitive Capacity Limitations**: The amount of potentially relevant geoscience information available exceeds our cognitive capacity to read it and detect subtle patterns. It is likely that this limits the discovery of surprising, insightful and valuable connections, supporting the use case for algorithms to assist geoscientists and MEMEX devices.

3. **Law of Inflated Expectations**: Myth and propaganda serving technological interests, may blind some of us, some of the time, to Law number #1



1 Inspired by Isaac Asimov 's Three Law's of Robotics
Alan Turing (1950) Imitation Game
Vannevar Bush (1945)
Ellul (1954) The Technological Society

Background - Cognitive Bias

"A compelling narrative fosters an illusion of inevitability" Daniel Kahneman

"People generally see what they look for, and hear what they listen for" Harper Lee

COGNITIVE BIAS

(Rose 2016)

Premature selection of theory
Personal hubris
Lack of perspective
Lack of imagination
Laziness
Excessive self-interest

BIG DATA SENTIMENT ANALYSIS FROM EXTERNAL LITERATURE / COMPANY ARCHIVES

CHALLENGE SURPISE US?

GroupThink is to be avoided ....

.... but being aware of independently stacked opinion in literature may challenge our own biases

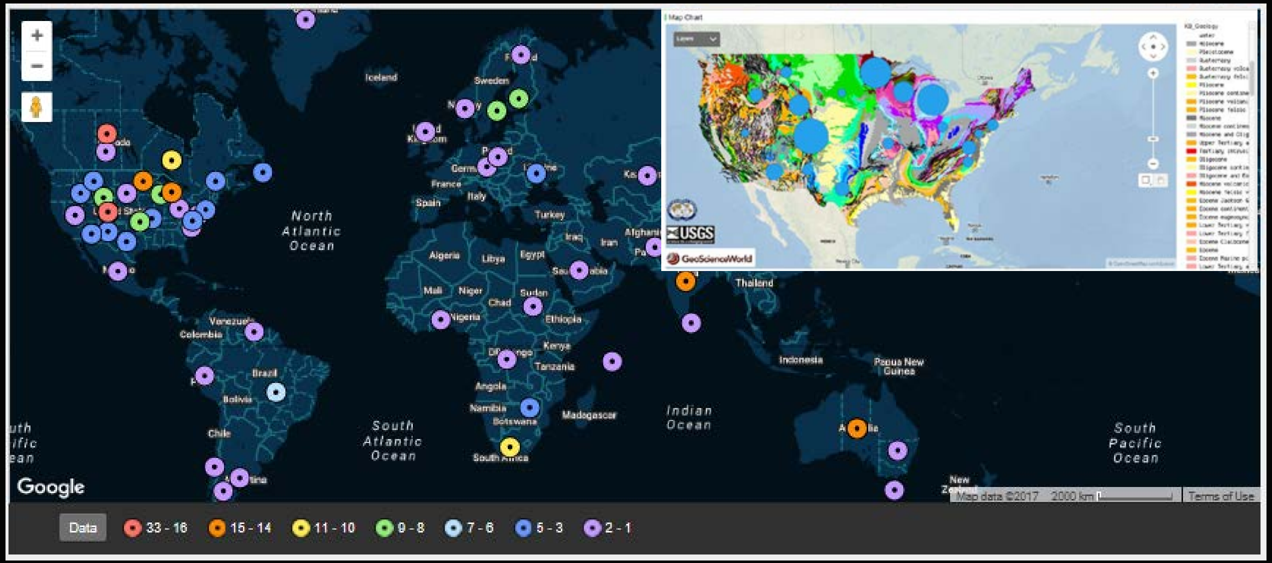
Text and Data Mining – Rule Based

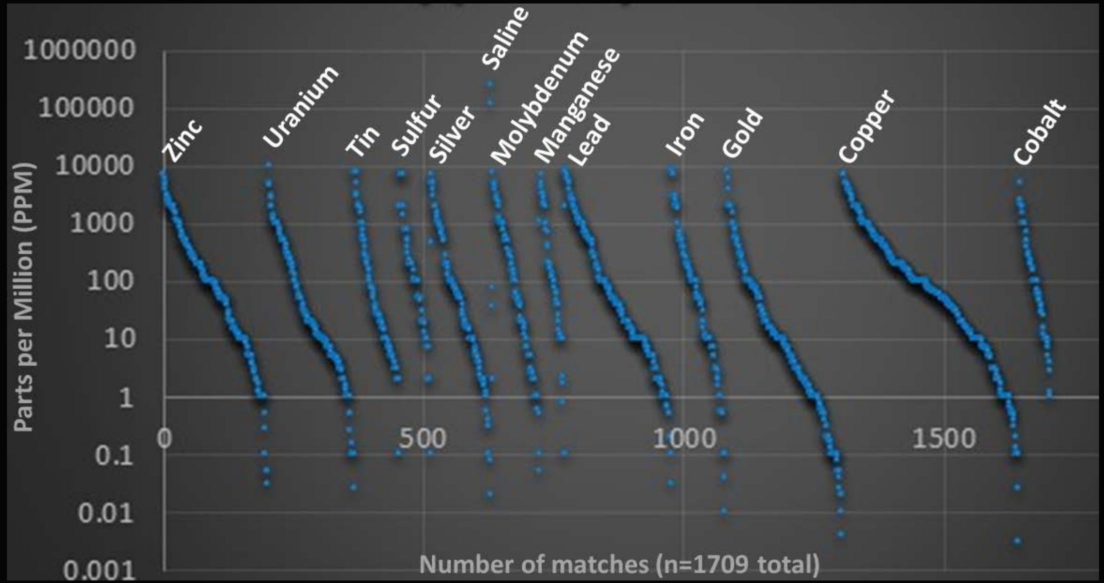
Spatializing entities/concepts and associations
e.g. ‘mentions’ of Pre-Cambrian

C shows histograms of δC analyses of anthraxolite and organic carbon in Precambrian sedimentary rocks, graphite in noncalcareous Precambrian schists and gneisses, and vein graphites from Ceylon, Montana, and New Hampshire. From these data, it is clear that

‘Extracting integer and float data from unstructured text
e.g. ppm is an association with a chemical element

```
Monroe,where a ridge on the east meets the plain of Sevier River.
It is discharging about 30 gallons minute of water at temperatures
range from per that 135ø to 146ø F; an analysis indicates that it
contains 0.41 ppm of manganese.
```





Cleverley (2017) Data courtesy of the Society of Economic Geology via GeoscienceWorld

Unsupervised machine learning: Word Vectors

Question: Which is the most similar play to the ‘Green Formation’?

	Word count of co-occurrences		
Play	**Reservoir**	**Source Rock**	**Trap**
Green Formation			
Beano Formation			
Nemo Formation			
…			

Answer: Nemo is the most similar play based on latent patterns in text

	Word count of co-occurrences			
Play	**shoreface**	**well rounded**	**fractured**	**…**
Green Formation: Reservoir	5	3	2	
Beano Formation: Reservoir	0	1	2	
Nemo Formation: Reservoir	1	4	3	
…				

$$\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Green Formation v Beano Formation = (5*0+3*1+0*5+2*2)/((SQRT(5*5+3*3+0*0+2*2)x(SQRT(0*0+1*1+5*5+2*2)) = 0.2075

Green Formation v Nemo Formation = (5*1+3*4+0*0+2*3)/((SQRT(5*5+3*3+0*0+2*2)x(SQRT(1*1+4*4+0*0+3*3)) = 0.7333

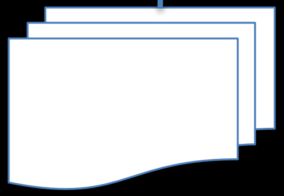
Cleverley, P.H. (2016) Machine Learning in Text for Geoscientists.
Digital Energy and British Computer Society, May 2016, Aberdeen, UK: https://paulhcleverley.com/2016/05/

Unsupervised machine learning: analogues

I need an analogue for.. play 23c Ask

Analogue	Overall	Reservoir	Source	Trap
Play 47	82%	91%	87%	63%
Play 12	80%	79%	82%	45%
Play 45a	65%	72%	15%	75%
Play 3	61%	91%	15%	25%
Play 102	58%	47%	20%	33%
Play 56	54%	34%	39%	62%
Play 32	51%	81%	18%	15%
…				

Co-occurrence Algorithms



Text Corpus (Documents)

To date I have only tested concepts with geoscientists on Lithostratigraphic Units not the whole play concept. Promising results:

“I input the xxx Formation that I studied in Tunisia and it returned a lateral equivalent (in Libya) that I had not come across before.“

Geologist, Multi-National Oil and Gas Company

Lexicons / Natural Language Processing (NLP) Entity Extraction/Noun Phrases
Concept Associative Extraction
Positive, PointWise Mutual Information Measure (PMI), SVM
DSM, Text Embeddings (Neural Network Word2Vec Mikolov *et al.* (2013); Řehůřek (2014))

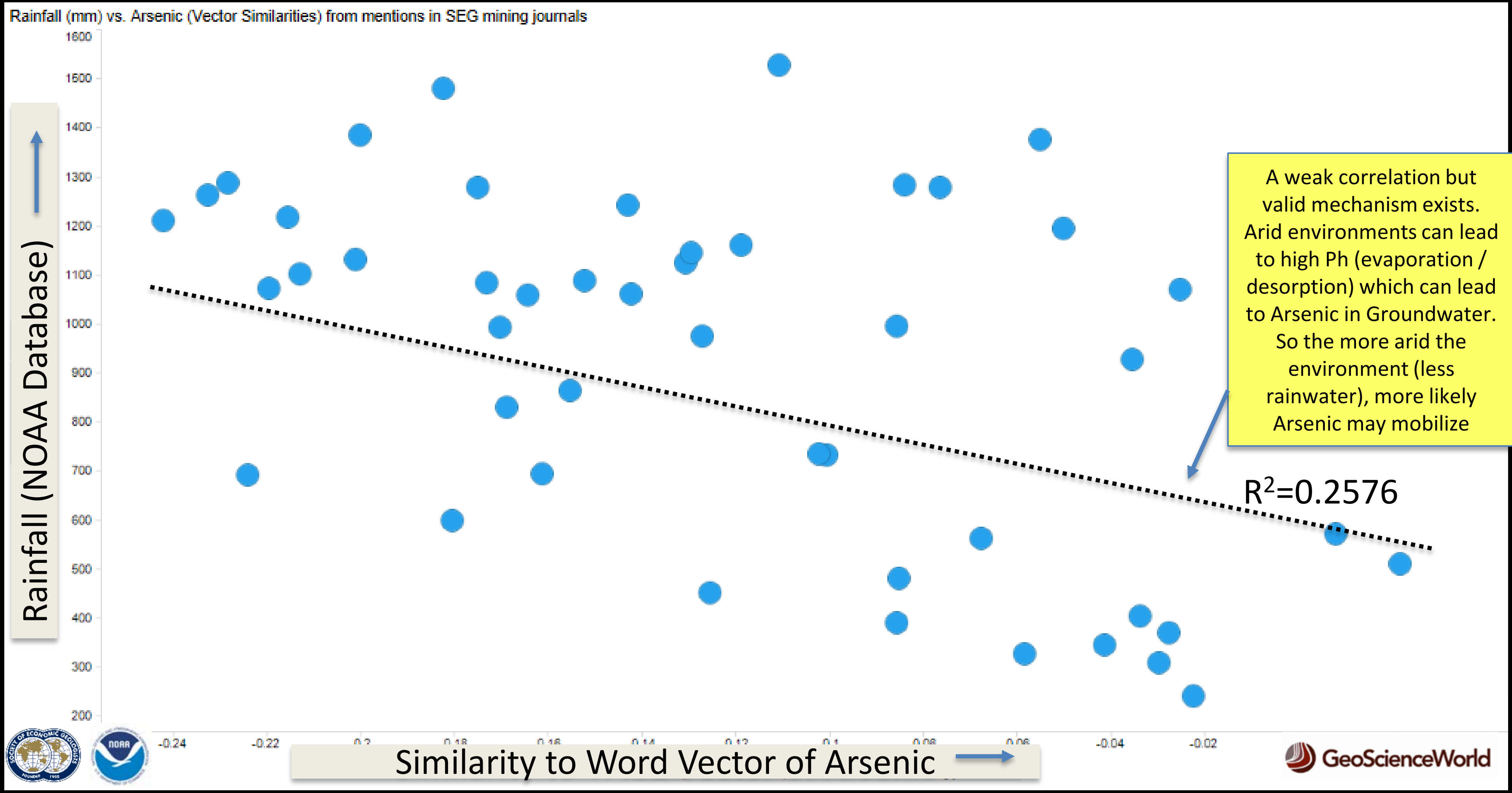
ROBERT GORDON UNIVERSITY ABERDEEN

Hypothesis Tester (Economic Geology Example)

US States

Y-Axis = annual rainfall data from NOAA

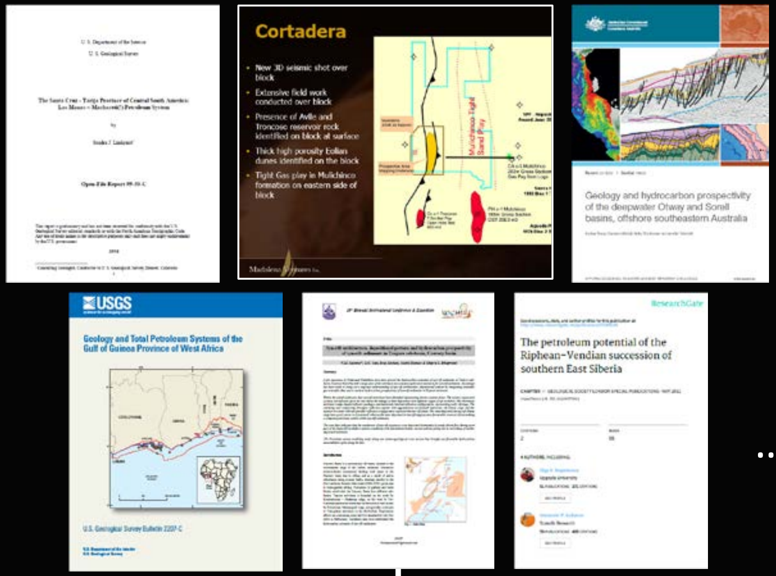
X-Axis = similarity of word vectors of all US States in SEG text (6,000 papers) to the word vector of Arsenic in SEG text



Cleverley, P.H. (2017) Text Analytics meets Geoscience. New Mexico, April 2017

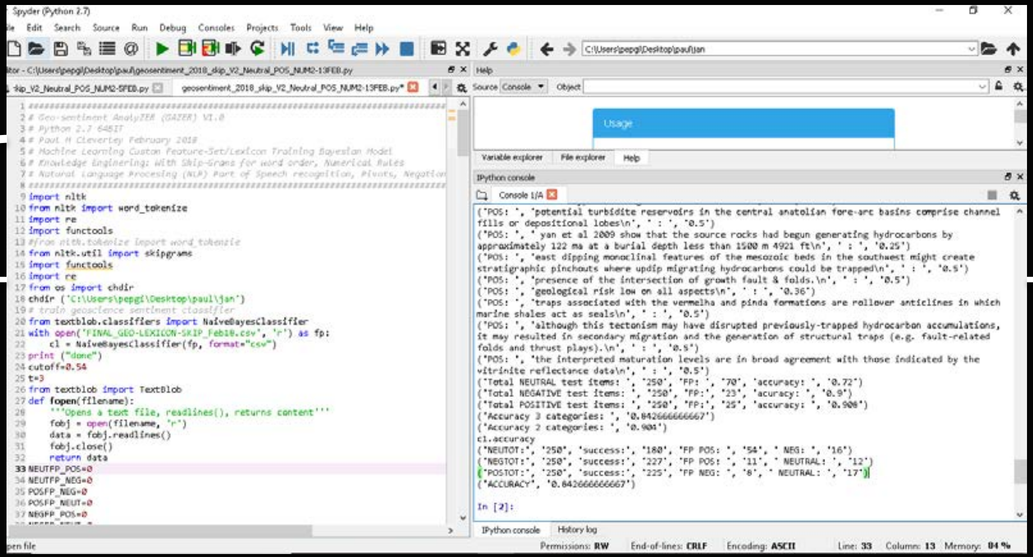
ROBERT GORDON UNIVERSITY ABERDEEN

Supervised Machine Learning: Sentiment Analysis



...

Labelled Sentences and Labelled Lexicons

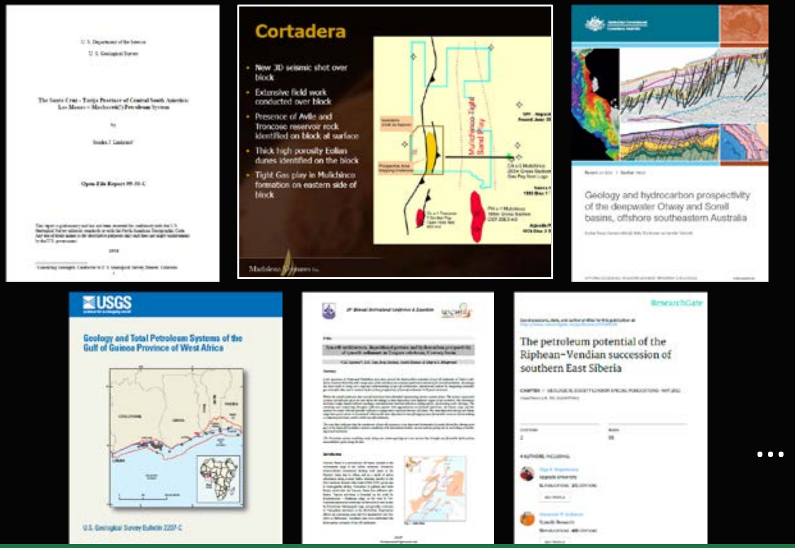


Python Code – Geoscience Aware Sentiment Analyzer (GAZER)

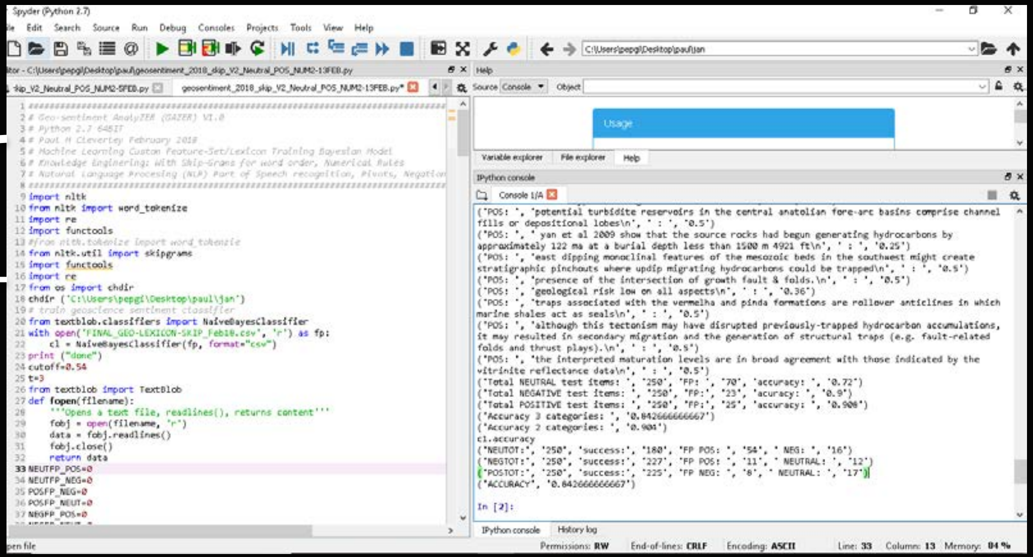
Small sample 150 Public Domain PDF Oil & Gas Reports - 5 Million+ words → Convert to ASCII text → Identify Petroleum System "Mentions" → 2,500+ Sentences → Retired Geologist's Label → MACHINE CLASSIFIER SENTIMENT

Agreed with each other 91.6% of the time

Supervised Machine Learning: Sentiment Analysis



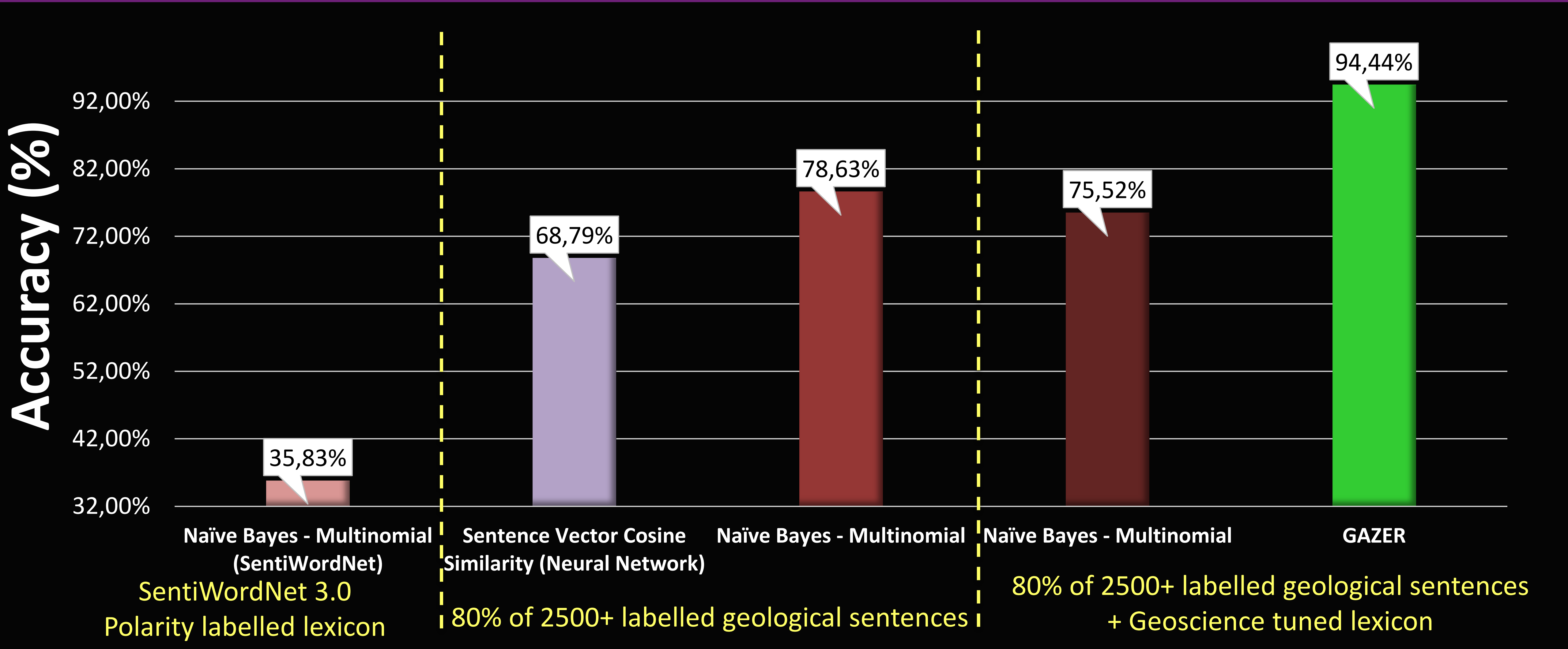
Labelled Sentences and Labelled Lexicons



TEXT SENTENCE AUTOMATICALLY EXTRACTED FROM DOCUMENTS	POLARITY
characterization of fluvial and aeolian reservoirs	Neutral
source rocks are thought to have expelled oil after trap formation	Positive
reservoir sands are not a problem in this area	Positive
..shows several attractive fault related traps but sedimentological and stratigraphic studies indicate that there would be poor reservoir characteristics..	Negative
the ro values between 0.5 and 0.7% indicate low source-rock grade	Negative
consequently, a key subsequent step in exploration was to establish the presence of source rocks in the basin	Neutral

ROBERT GORDON UNIVERSITY ABERDEEN

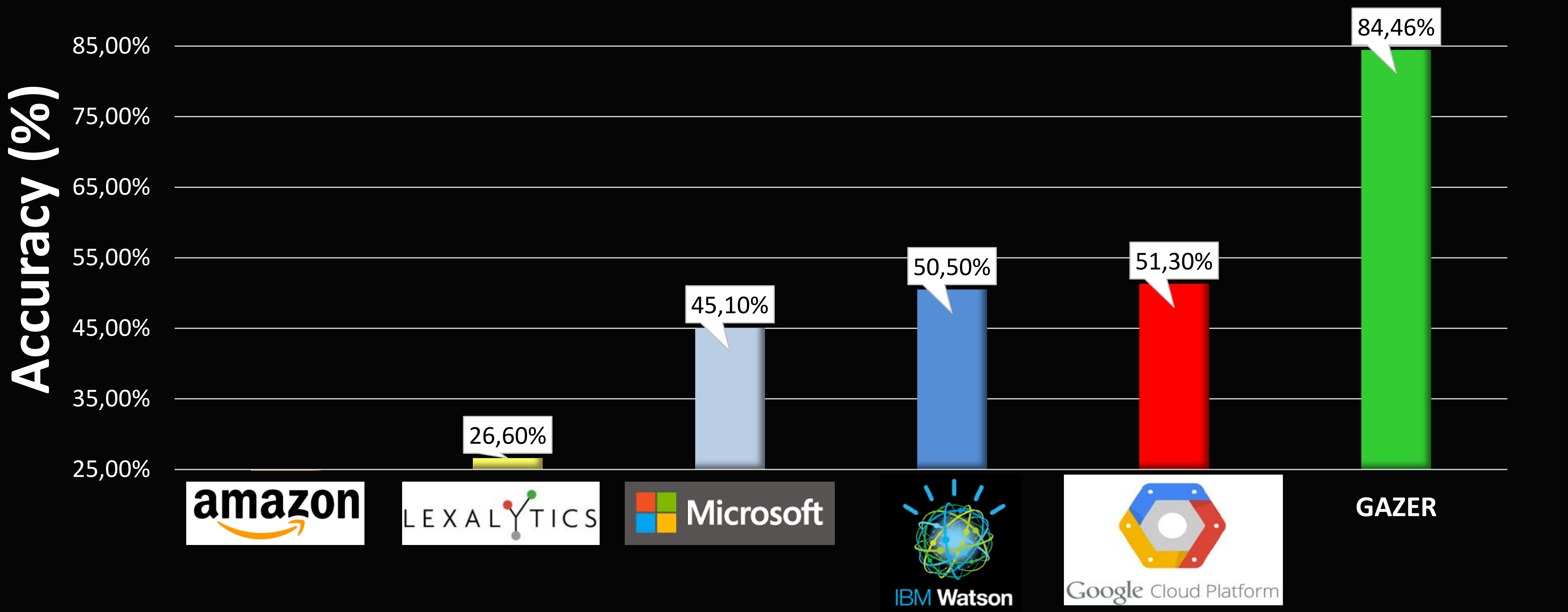
Results – Accuracy Comparison 2 Categories POS v NEG



- Comparing too State-of-the-art in the literature for generic sentiment analysis: Sentence Vector (Le & Mikolov 2014) gave 92.6% with 25,000 Move Reviews as a training set

ROBERT GORDON UNIVERSITY ABERDEEN

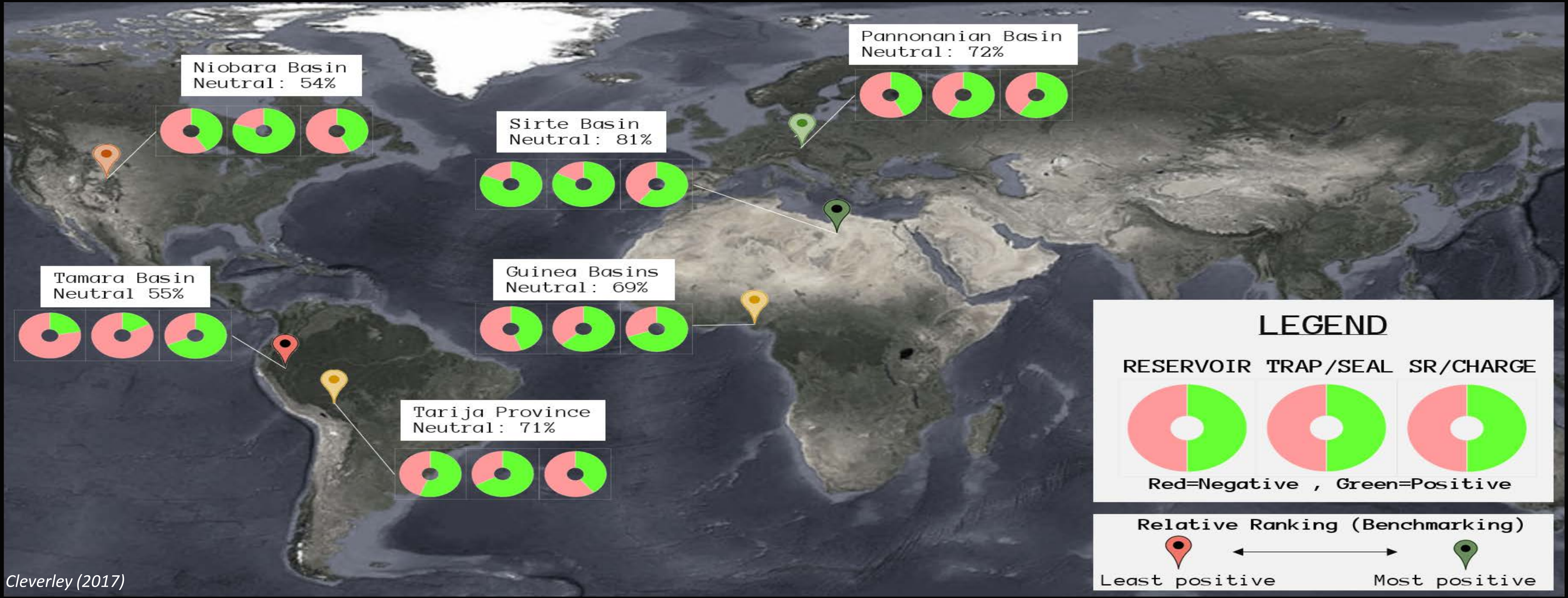
Results – Using 750 Test Set 3 Cat. (POS-NEG-NEUT)



Published data sets
750 labelled sentences (POS, NEG, NEUT)
for test benchmarking classifier performance on Github:
https://github.com/phcleverley/Geoscience-Sentiment-Research

ROBERT GORDON UNIVERSITY ABERDEEN

Data driven insights



Cleverley (2017)

"…a really nice way to capture literature without reading it. It would be nice to run this method on all the AAPG publications of a few separate basins and see if the graphs reflect our basic understanding of these basins. This could become a very powerful method in understanding and visualizing the current state of knowledge." Exploration Geologist (December 2017)

ROBERT GORDON UNIVERSITY ABERDEEN

Summary and future work

- Extracting text & numerical data from documents is only half the story..
- Algorithms can approach ‘geoscientist-like’ accuracy on narrow classification tasks like sentiment analysis, but can process exponentially greater volumes of info.
- Finding analogues is a ‘similarity’ challenge. Identifying what similarity dimensions and characteristics appeal to geoscientists (and in which contexts and why) is an area for further research.

Cleverley & Burnett (2014) Study of 53 geoscientists using various stimulants on touchscreens

Email: p.h.cleverley@rgu.ac.uk
Blog: www.paulhcleverley.com