

Text Analytics Hacks - a Cheat Sheet for Insightful Text Analytics

Andrew Marritt | Founder & CEO

andrew.marritt@organizationview.com | +41 78 905 0775

What we do - making text a solved problem for HR

OrganizationView was one of the first People Analytics firms. We believe that the most effective decisions are made when they're supported by both quantitative and qualitative data.

**Custom 'Inductive'
models**

Standard Text Models

Text data advisory

Insight from processed text data

Lots of people analytics content in our weekly newsletter: www.EmpiricalHR.com

Text analysis as quick, often exploratory analysis

Quantitative Analysis

Explains what happens / could happen (when, to what etc.)

Time consuming

Do you have the right and enough data?

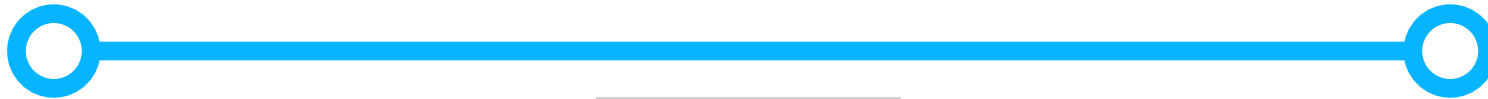
Qualitative Analysis

Explains why something is happening

Time consuming, but don't need as many subjects

Effective at driving action

Does it scale?



Text Analysis

Can indicate both what & why

Quick & 'cheap'

Doesn't need a pre-built 'hypothesis'

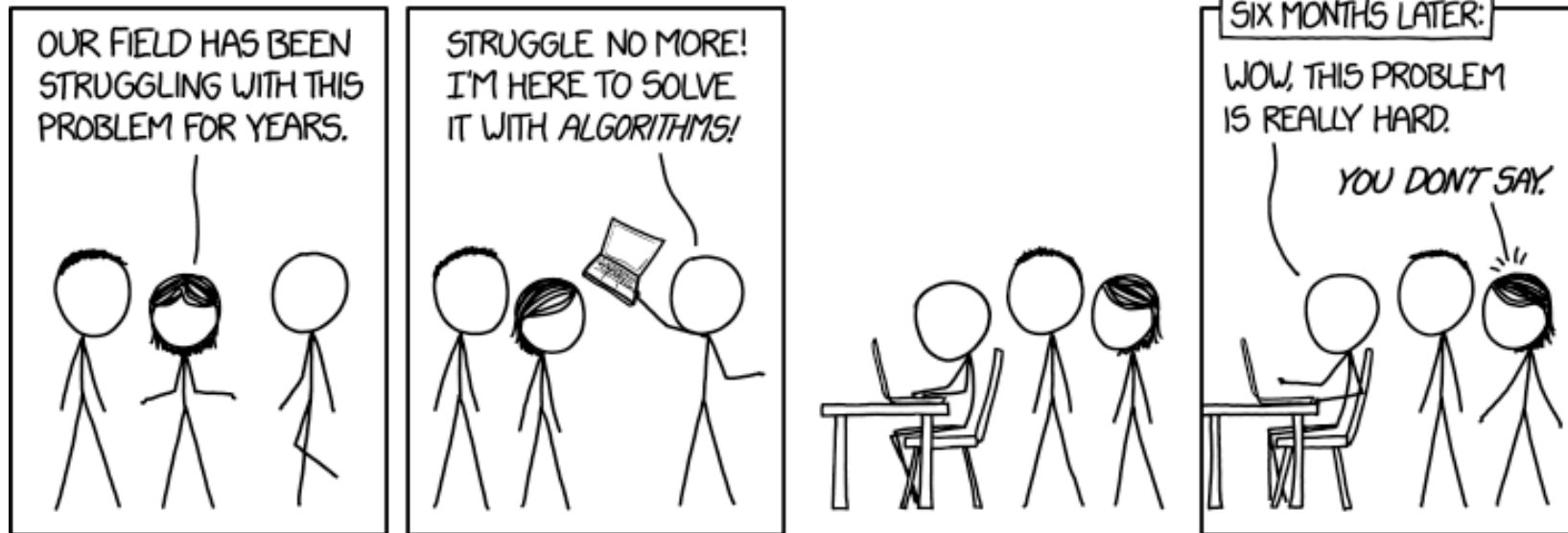
Can usually find existing data sources

Good at showing linkages

Short distance to detail

There are no easy solutions - everything is in the long tail

But we're getting good at solving specific problems



WHY TEXT ANALYSIS IS DIFFICULT

Challenges with text 1: Vast number of features

A recent, medium-sized dataset

Number of answers	28,500
Number of words	1,145,537
Number of unique words	18,529

Challenges with text 2: Single words aren't good enough

Keyword searching is never good enough.

Better Management != Performance Management
Data Management
Financial Management
Supply Chain Management
Document management
Time management

Balancing inaccuracy of words by n-grams adds complexity

A recent, medium-sized dataset

Number of answers	28,500
Number of words	1,145,537
Number of unique words (1-gram)	18,529
2-grams	218,097
3-grams	550,559
4-grams	752,914
5-grams	810,386

Sometimes they don't even mention the words!

Context is everything!

“lol, a raise!”

Correctly identified by our algorithms as someone wanting a pay increase

NEED TO BUILD A TRAINING SET

Text classification requires manually created training data

There is no getting away from needing good training data

Most predictive models in HR:

There is access to a dataset which shows what you want to predict that the model can learn from.

<i>Label</i>	<i>Gender</i>	<i>Age</i>	<i>Location</i>
1	M	32	DE
0	F	25	UK
0	M	58	ES

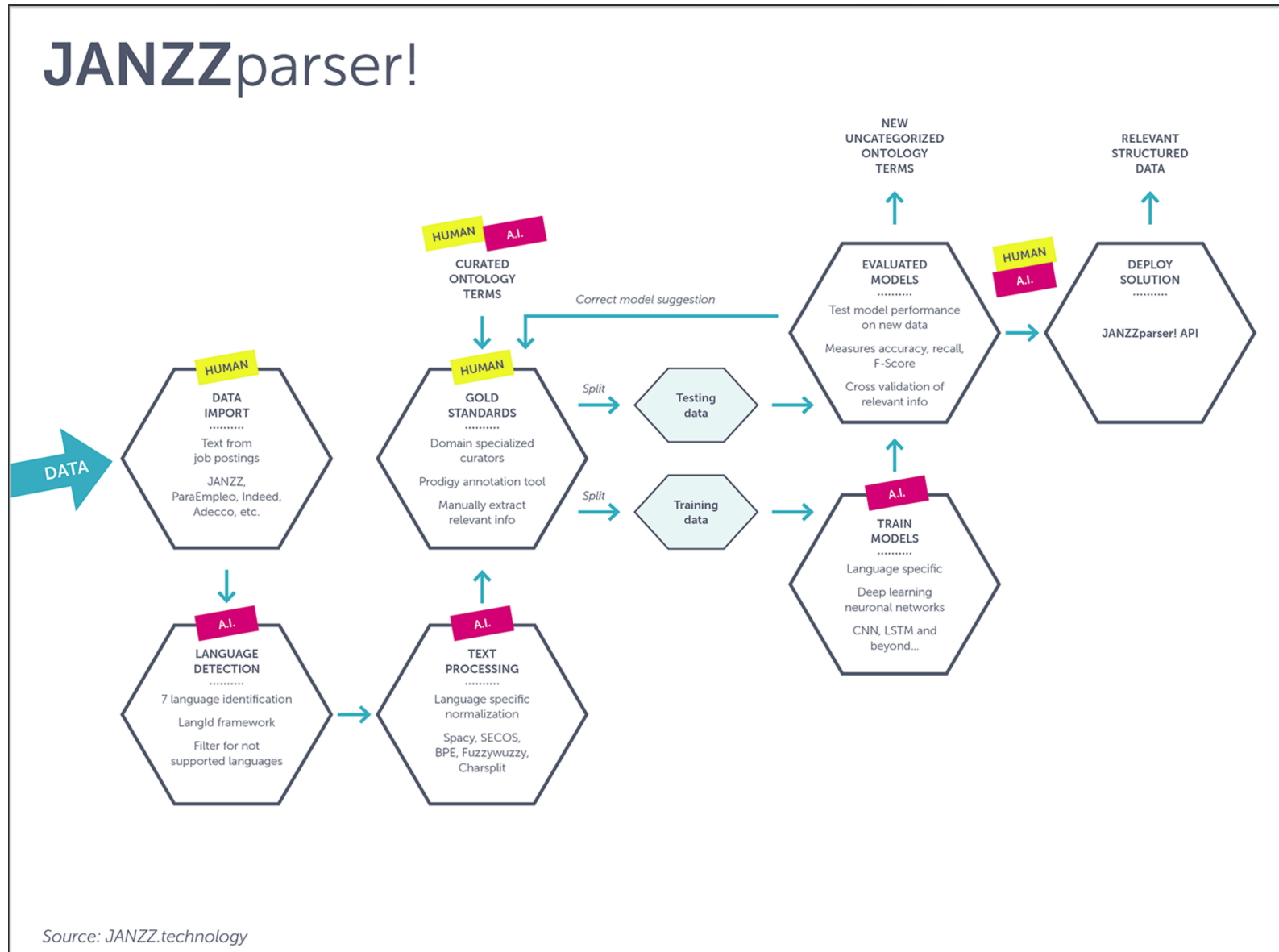
Text classification

No model exists - you have to create one!

<i>Theme</i>	<i>Sentence</i>
Pay	Can you increase my wages
Training	After the grad scheme there is no training available
Manager	My manager never provides training
Training	My manager never provides training

The manual labelling costs can be enormous!

Fellow Swiss HR text firm Janzz have 20 people manually coding data for job / cv matching. They've been doing this for 10 years.



It's not just hard work, it requires highly-skilled teams

WIRED

Google's Hand-Fed AI Now Gives Answers, Not Just Search Results

Deep learning is changing how Google's search engine works. But its new-found efficiency takes a lot of painstaking human work behind the scenes.

“To train Google’s artificial Q&A brain, Orr and company also use old news stories, where machines start to see how headlines serve as short summaries of the longer articles that follow. But for now, **the company still needs its team of PhD linguists.**

They not only demonstrate sentence compression, but actually label parts of speech in ways that help neural nets understand how human language works. **Spanning about 100 PhD linguists across the globe,** the Pygmalion team produces what Orr calls “the gold data,” while the news stories are the “silver.””

BEING SPECIFIC

Challenges in text 5: False positives kill your credibility

To get buy-in you need to be good. People are more forgiving to not understanding than to getting it wrong.

The Google Assistant can help you get things done over the phone



The system only works when it has been trained in extremely narrow contexts, or “domains”, ... since the range of the language used is limited and it is easier to predict what people will say next.

0:07 / 1:27



The context of the question is vital

It's not going to be possible to develop a good universal text model anytime soon

How you interpret a sentence depends on the question that was asked.

Respondents answer assuming you have the context of the question. Often the answer won't mention the theme of the question explicitly

What insight is important to the decision maker depends on the question

If themes are to convey the most meaning they need to be contextual on the decision that they will inform.

Context of the organization is highly beneficial

Employees will often use names rather than descriptions. They assume you know these.

Bangalore

The offshoring centre

Edinburgh

The headquarters

Jane

The CEO

‘Ask Anita’

The internal chatbot

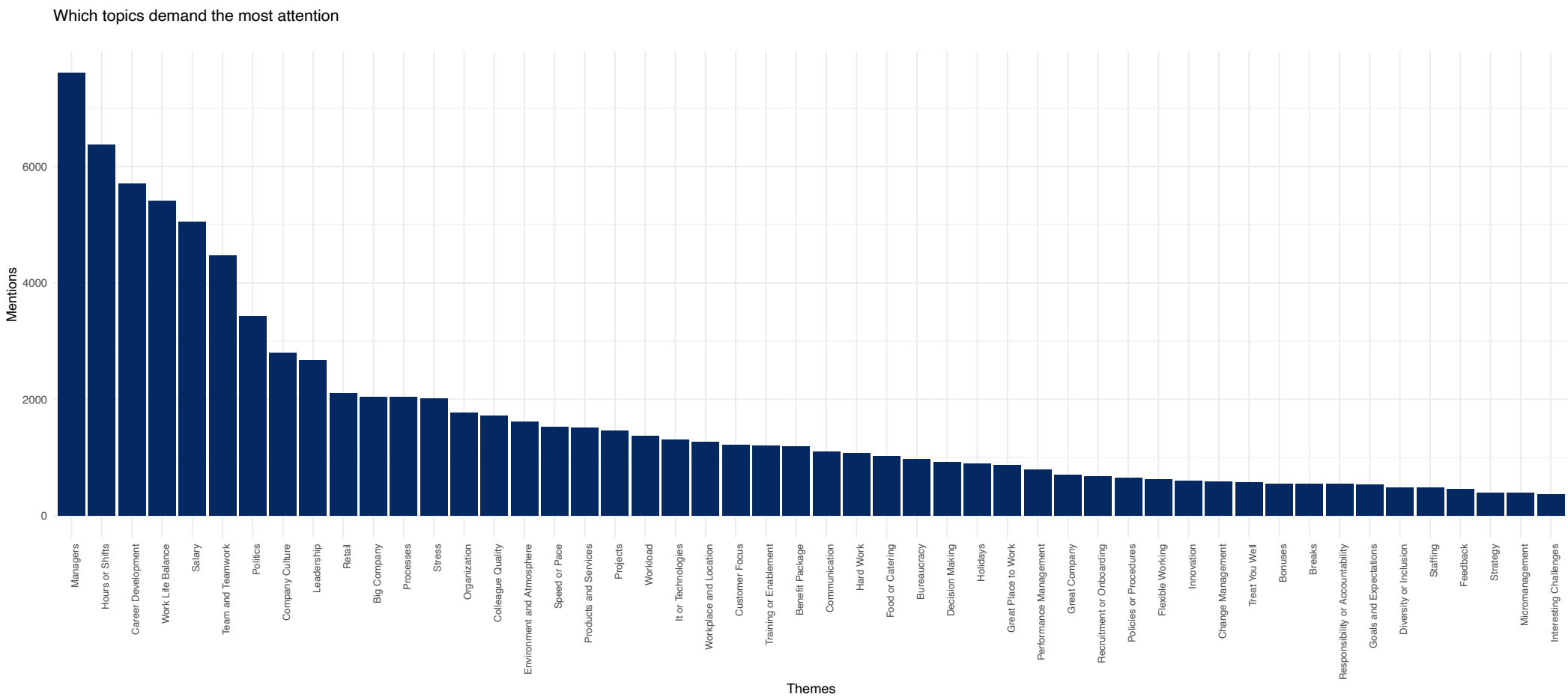
Eagle

The finance transformation programme

TRAINING DATA: VOLUME vs VARIETY

Challenges with text 3: Everything is in the long tail

The easy examples are easy; what will determine success are the rare ones!



Need to identify and support the algorithm where it's struggling

Understanding your algorithm and where it's not converging is vital to getting really high performance.

How do we separate....	from this
my workplace (the place I work)	Workplace , the software from Facebook
an employee's benefits package	how an employee benefits from....
more pay	pay more attention
better health	better health insurance
Moral (most common misspelling of Morale)	Moral (ethics)

HOW TO COLLECT BETTER TEXT

**Objective is to minimise false positives
(even if this means more false negatives)**

Executives are more
'understanding' of
'don't know' than
getting it wrong

Easier to review and
'fix' false negatives

It's easier to measure
model performance
in absence of 'gold
standard' test set

A well written question makes a big difference in coding accuracy

Questions should be open enough to encourage employees to speak....

but not so open that they deviate off what you care about.

Sentiment analysis struggles to classify text => Ask a positive and negative question

“What do you think works well in terms of our innovation approaches?”

“What could we do to improve innovation at COMPANY?”

Other items and questions on surveys make a big difference in the text.

If you’ve asked lots about topic X, the respondents will be thinking about topic X

Improving survey design is the number one thing you can do to improve your text analysis!

Understanding the answer by knowing the question

“For the amount of effort we do we should make more.”

Q1: What can we do to improve your working life here? (improve pay)

Q2: What ideas do you have to improve this factory? (improve productivity)

EMBEDDINGS vs STATISTICAL APPROACHES

Why the move towards embeddings

Word embeddings first appeared in ~2012. Now most high quality applications

- If we can reduce words to numbers it will be easier for computers to do calculations
- In 'Bag of words' we counted how often a word occurred, often in the presence of another word
- We still have a problem that the number of words (and misspellings etc.) is huge
- "Words are characterized by the company they keep"

Embeddings address these ideas by dramatically reducing dimensionality

What is special about Embeddings?

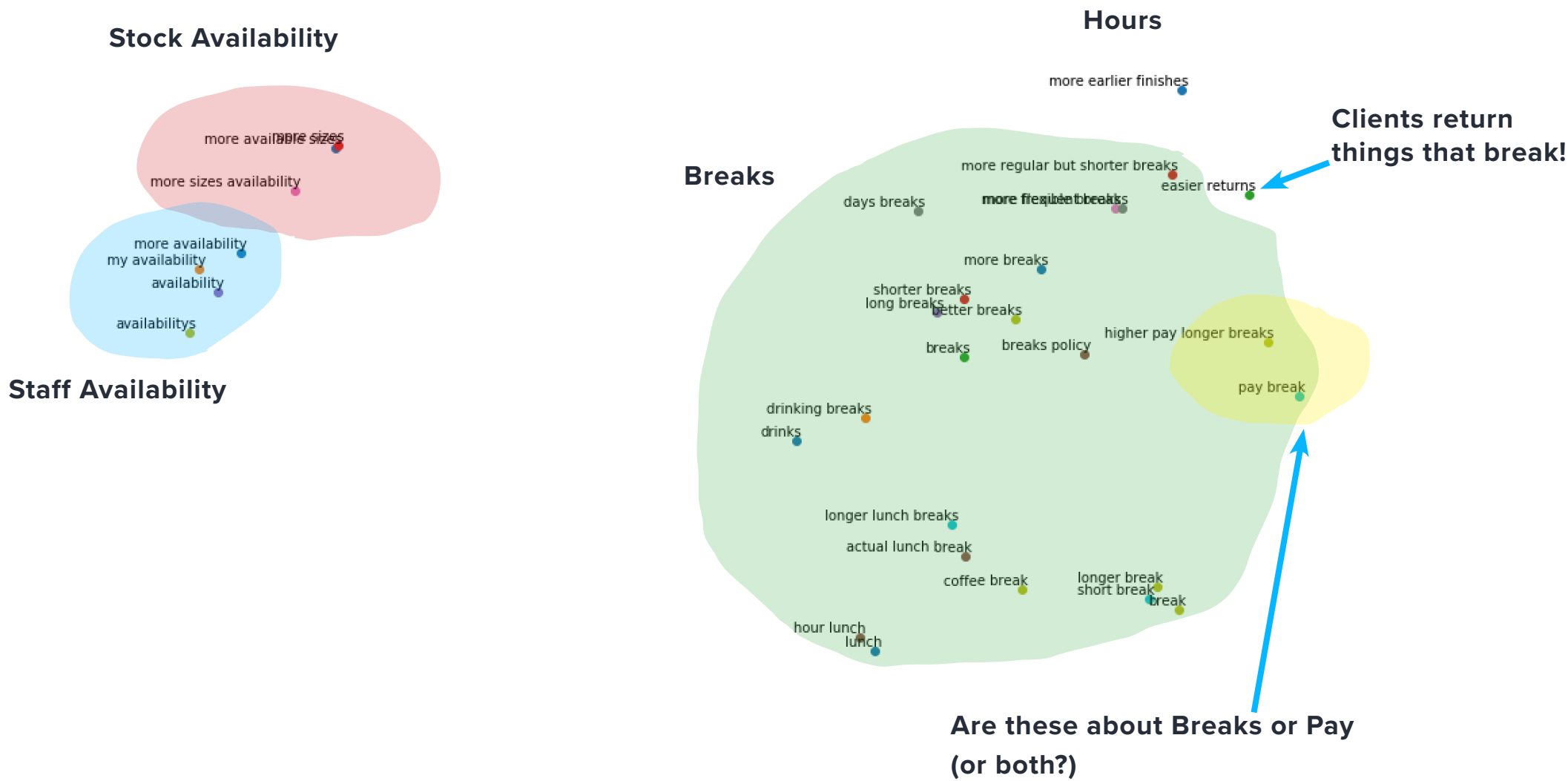
The process of creating embeddings creates semantic information. Depending on the algorithm it can also correct for misspellings etc.

Embeddings are matrices built from how frequently terms co-occur

- Two words that are associated are close to each other
- Distances can be calculated to show 'similarity'
- We can easily cluster words via association to create themes
- The embeddings can be used in machine learning models
- Most matrix calculation techniques are valid
- Embeddings can be at many levels from words to sentences

Embeddings group semantically similar concepts together

In reality we use many dimensions. The challenge is defining the boundaries.



Embeddings can do amazing things

Facebook Research showed you could get good translation by linking only a few pairs of words

facebook research

Search



Research Areas

Publications

People

Academic Programs

Downloads & Projects

Careers

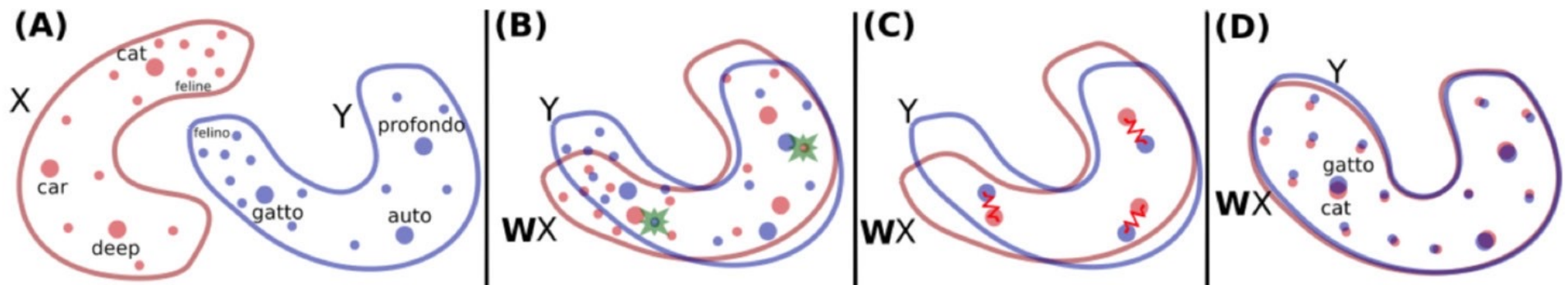
Blog

April 30, 2018

Word Translation Without Parallel Data

International Conference on Learning Representations (ICLR)

By: Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou



SENTIMENT ANALYSIS

Sentiment analysis isn't very good

The current state of the art is about 65% accuracy.

Try the API



You always surprise me with your mediocracy

ANALYZE

[See supported languages](#)

Entities	Sentiment	Syntax	Categories
----------	-----------	--------	------------

Document & Sentence Level Sentiment

Entire Document

You always surprise me with your mediocracy

Score Magnitude

0.3

0.3

0.3

0.3

Score Range

-1.0 — -0.25

-0.25 — 0.25

0.25 — 1.0

If sentiment isn't accurate how accurate is emotion?

Document-level

Tones

Anger ☐ Fear ☐ Joy ☐ Sadness ☒ Analytical ☒ Confident ☐ Tentative ☐

[View JSON](#)

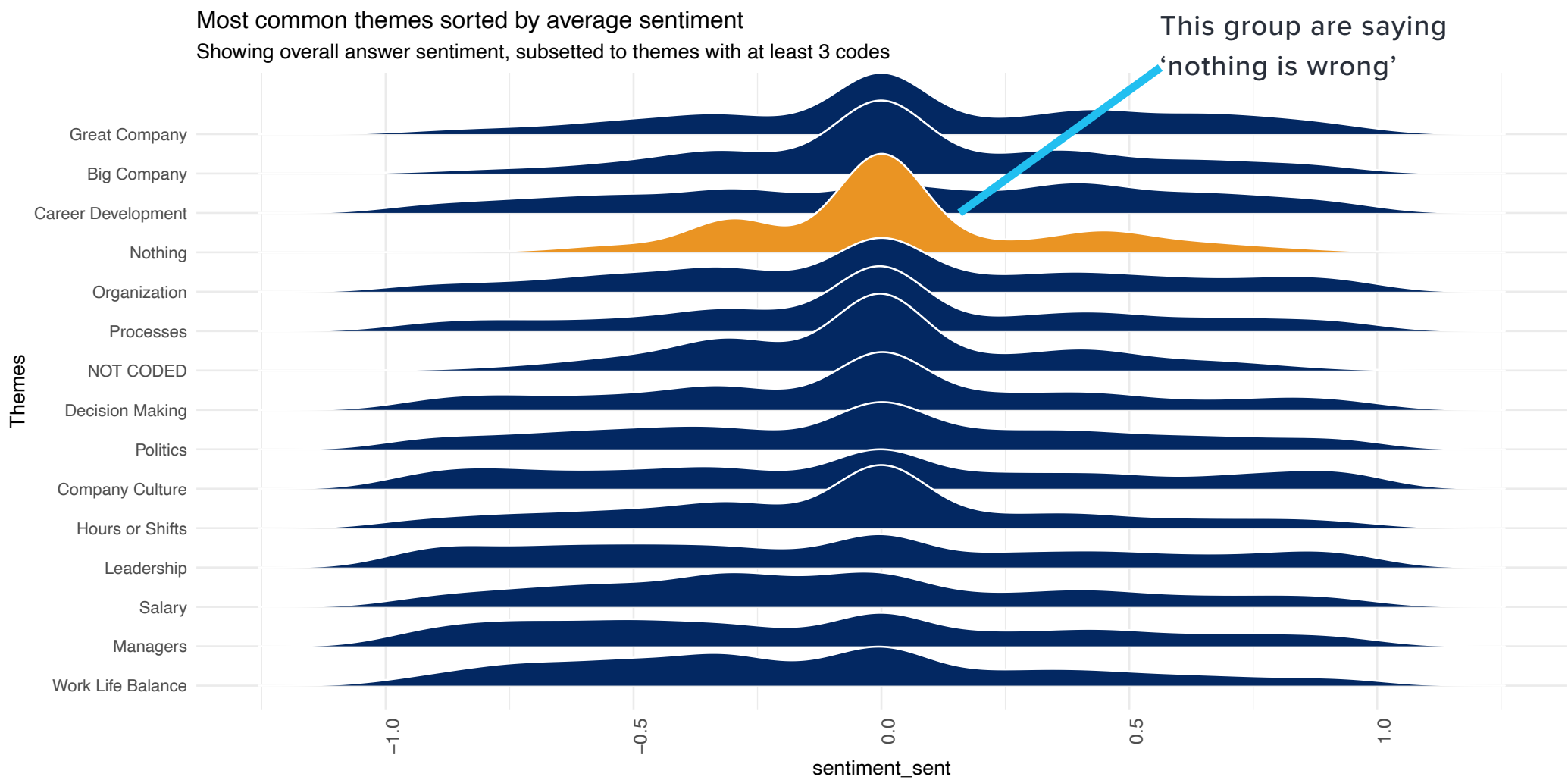
Sentence-level

Identify sentences with stronger tones in context or sorted by score. Highlighted sentences indicate the likelihood of a tone present. If more than one tone is present, the stronger one is shown. Click on a sentence to see a breakdown of all tones.

Tones	In context
Analytical	
Confident	
Sadness	Sadness: Indicates a feeling of loss and disadvantage. When a person can be observed to be quiet, less energetic and withdrawn, it may be inferred that sadness exists. < .5 .5 - .75 > .75 NoneStrong Also it not tare on customers. In addition the quality of store closing has gone down hill and the customers are returning a lot items because of that. Also working for 6 hours without a break is really had for a lot of staff. I love working for this firm but at times it's really stressful when communications are bad between staff and head office. If possible as we'll in delivery can they please send more 3 and 9s, I say this every week and nothing changes. Customers are compiling that we never have those sizes in store Because we get 1

Especially on negative questions sentiment is next-to-useless

Negative questions include ‘what are the cons of working here?’ or ‘what do we need to improve?’



ANALYSIS POST THEMES

Text analysis is often a means to an end

Always think back from what you want to achieve rather than think you need 'text analysis'.

Just coding the text

Limited use cases,
mostly around filtering

Coding then pattern recognition

What can we say about
the codes and who
mentions them?

Coding as feature generation

Text analysis to
understand other data

Temporal analysis of text

What is trending? How
does language evolve?

“Look here analysis”

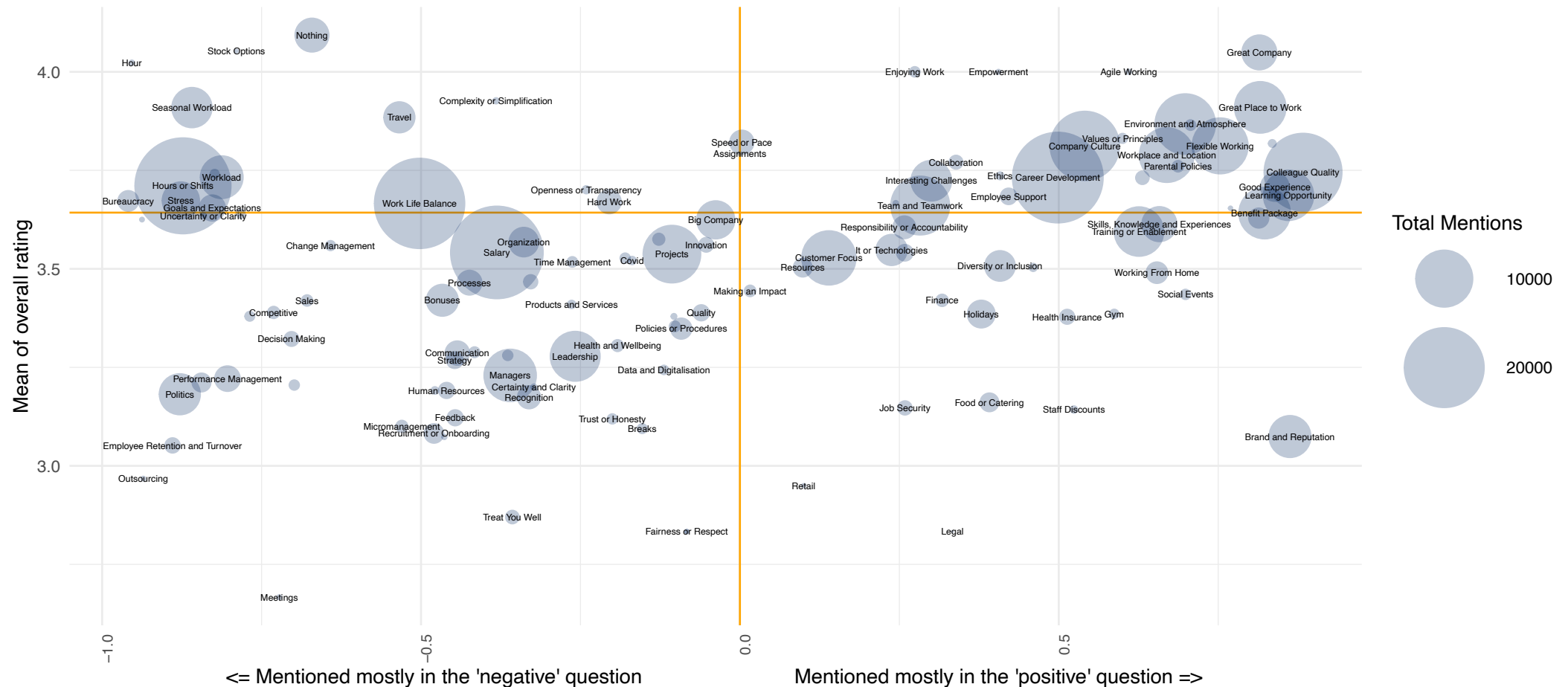
How to guide
executives to important
items of text

Coding in automation

Text analysis to route
queries or allow
automation

Using a key variable as an alternative to sentiment

Sentiment measured by the average overall rating of reviewers mentioning that theme and the balance between positive and negative mentions
Orange horizontal line shows mean overall rating (3.64)



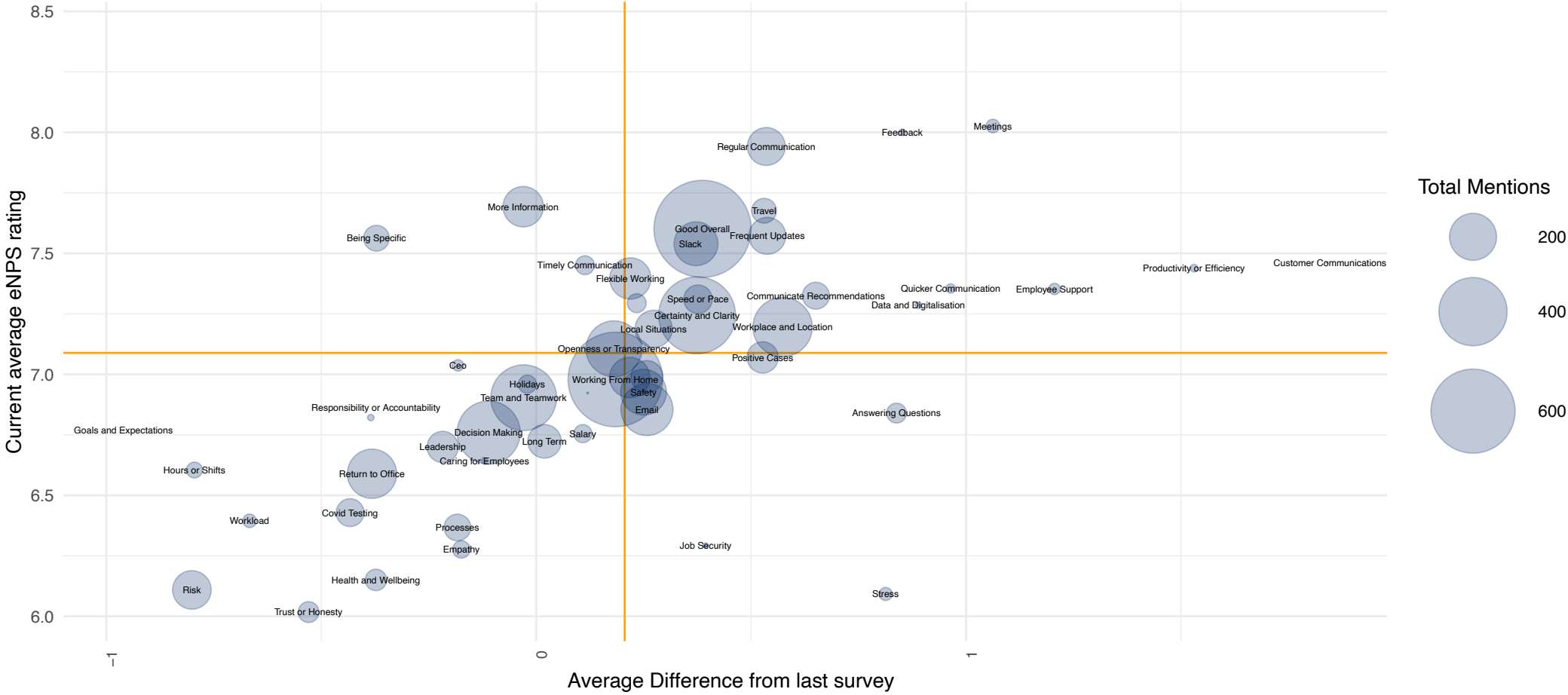
Understanding changes in perspectives

Longitudinal ‘panel’ data lets us explore topics associated with changes in NPS

How are the themes related to an individual's current and recent change in NPS?

Orange reference lines indicate the current mean score or change.

Showing only themes with at least 25 mentions



Co-occurrence shows patterns in ‘speech’

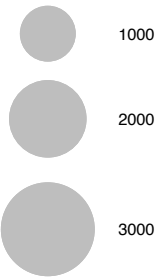
Network visualisation of topic cooccurrence

Showing relationships where the probability of them being strongly related is at least 98%

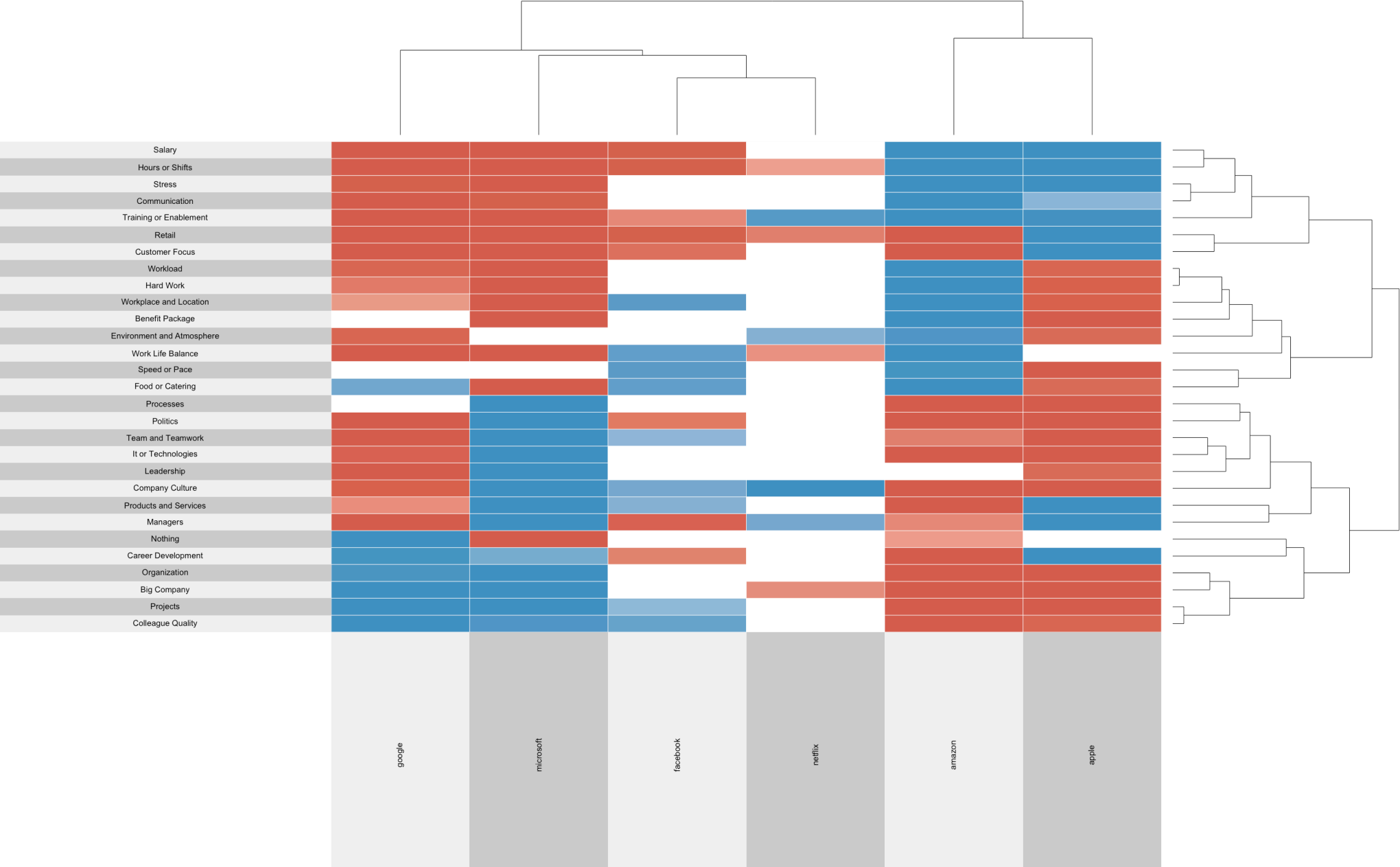
Community

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

Mentions



External reviews can enable benchmarking / competitive insight



TRANSFORMER-BASED APPROACHES

PAST, PRESENT, FUTURE

History of text analysis

3 major time periods in text analytics

**NLP / Programmed
Grammatical
Approaches**

Statistical Approaches

**Deep Learning / Word
Embeddings**

Using massive datasets and transformer based models computers are being able to learn language

[illegible]

GPT-3 required 499B tokens to train (all Wikipedia has 4B words)

GPT-3 has 175 billion parameters and would require 355 years and \$4,600,000 to train - even with the lowest priced GPU cloud on the market.

Tools that you can use

If you want to do this yourself you're going to need a scripting language

R	Python
Quanteda	NLTK
Text2Vec	CoreNLP
SpaCyR	SpaCy
TidyText	Gensim
Stringr	TextBlob
tm	Hugging Face

Parsing trees and Parts of Speech

Historically we struggled most with short answers without any grammatical structure!

Try the API

Text Analytics is a key part of People Analytics

ANALYZE

[See supported languages](#)

Entities

Sentiment

Syntax

Categories

☒ Dependency

☒ Parse Label

☒ Part of Speech

☒ Lemma

☒ Morphology

nn

Text

NOUN

number=SINGULAR

proper=PROPER

nsubj

Analytics

NOUN

number=SINGULAR

proper=PROPER

root

is

VERB

mood=INDICATIVE

number=SINGULAR

person=THIRD

tense=PRESENT

det

a

DET

amod

key

ADJ

attr

part

NOUN

number=SINGULAR

prep

of

ADP

nn

People

NOUN

number=PLURAL

proper=PROPER

pobj

Analytics

NOUN

number=SINGULAR

proper=PROPER



Do your employees write in sentences?

Preprocessing is a balance

Everytime in analysis when you aggregate or simplify you destroy potentially valuable information

“My system keeps crashing! his crashed yesterday, ours crashes daily”

Stemming

‘My system keep crash hi crash
yesterday, our crash daili’

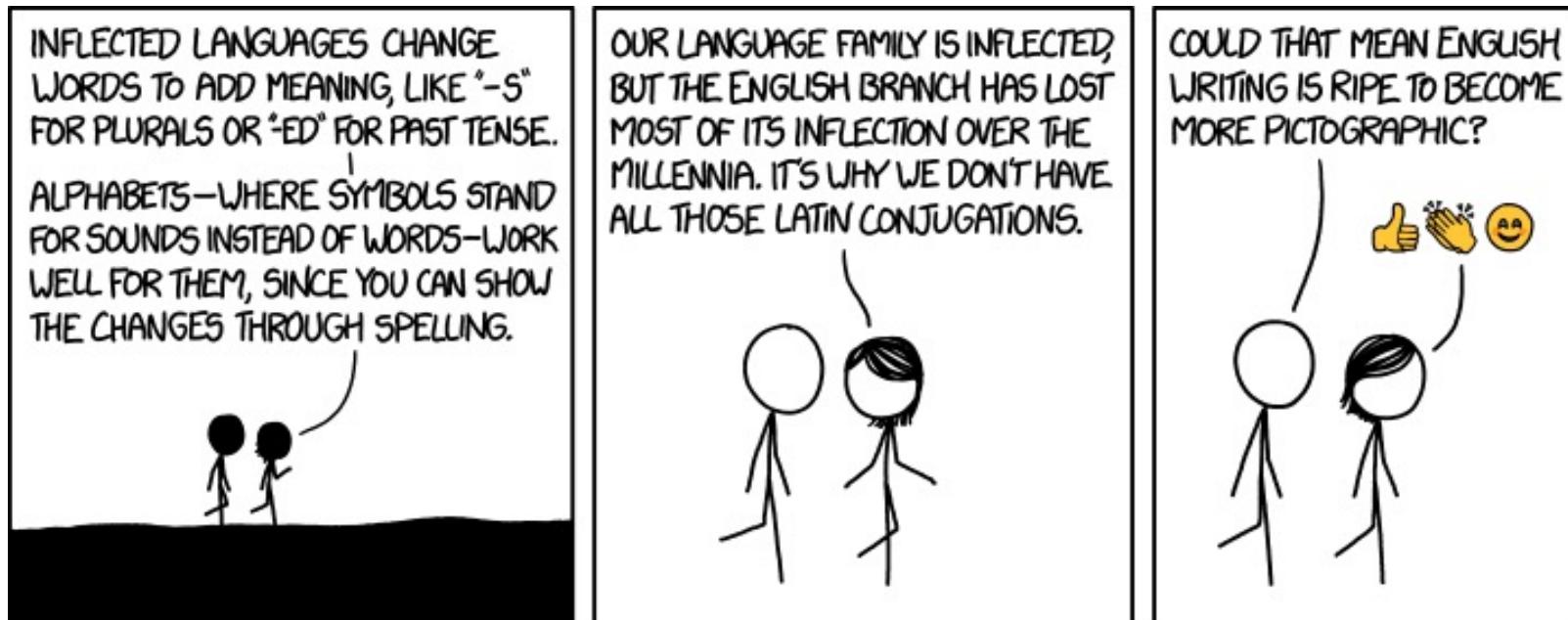
Lemmatization

‘My system keep crash ! his crash
yesterday , ours crash daily’

Communication isn't just words

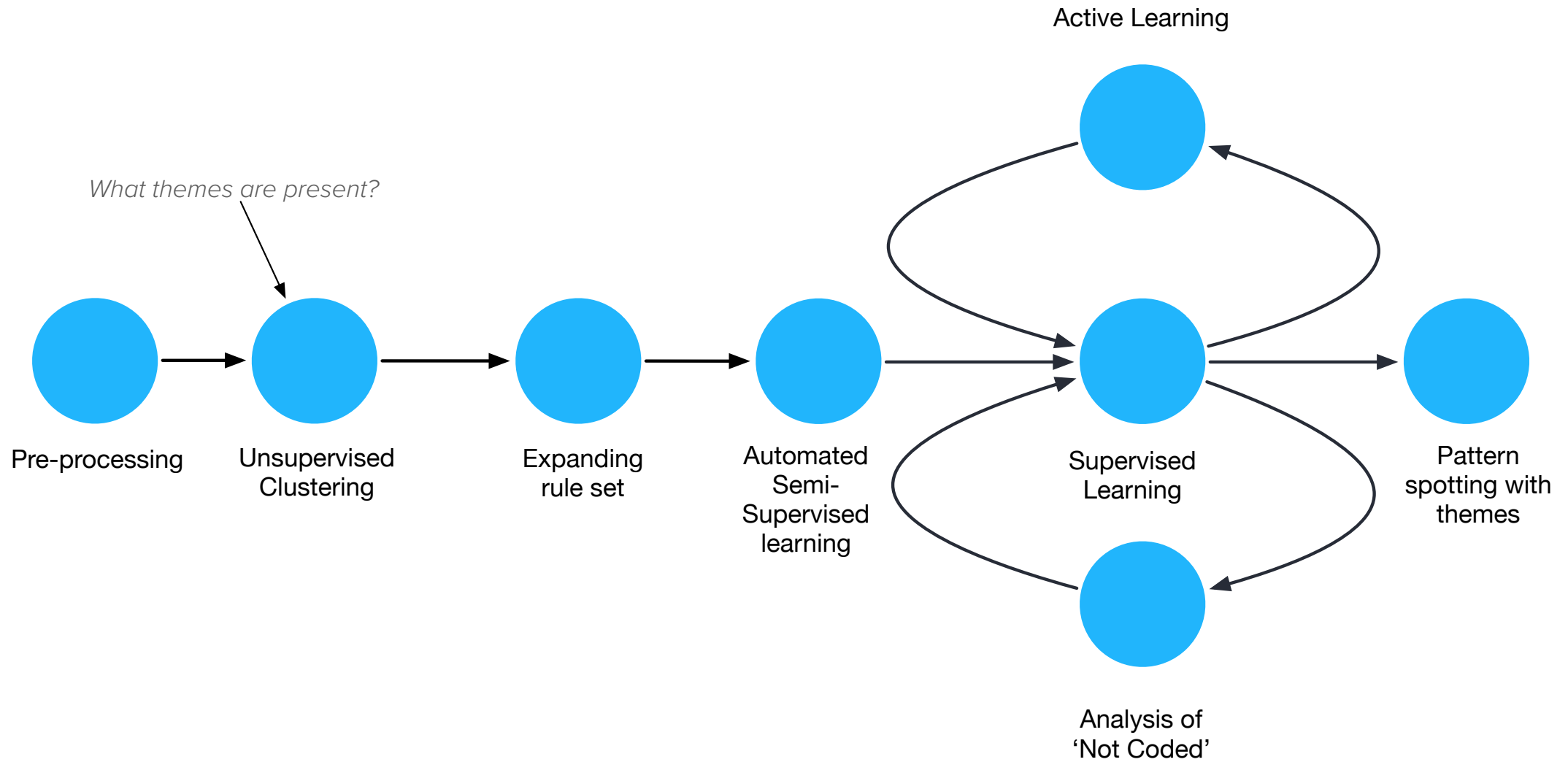
Emojis, text speak, acronyms.....

If you want to understand speech you might need to analyse non-speech parts (eg does a pause carry meaning?)



Putting it into practice

A high level overview of the Workometry text theming process



How we structure our team

Our key focus is on industrialising this process to be quicker, with lower errors and cheaper

Data Science team

Develops algorithms,
tests ideas, monitors
performance

Infrastructure team

Develops
WorkomemtryCI, our
internal coding platform

Coding and Curation team

Builds the models by
editing training data

Infrastructure can dramatically reduce training time

Our Coding Interface enables us to manually tune each theme independently

Sort			ALPHABETICALLY	BY MEDIUM	BY MEDIUM PERCENTAGE
Assignments	Last revised: 2021-04-05 06:05:34		Training examples: 1		
LOW: 0 MED: 1 HIGH: 0					
Enjoying Work	Last revised: 2021-04-05 06:01:12		Training examples: 5		
LOW: 3 MED: 3 HIGH: 0					
Commute	Last revised: 2021-04-09 15:18:50		Training examples: 18		
LOW: 6 MED: 4 HIGH: 2					
Openness Or Transparency	Last revised: 2021-04-05 06:05:30		Training examples: 5		
LOW: 2 MED: 1 HIGH: 0					
Certainty And Clarity	Last revised: 2021-04-05 06:14:45		Training examples: 31		
LOW: 7 MED: 3 HIGH: 1					
Data And Digitalisation	Last revised: 2021-04-13 10:36:58		Training examples: 30		
LOW: 12 MED: 4 HIGH: 0					
Meetings	Last revised: 2021-04-05 06:02:51		Training examples: 9		
LOW: 9 MED: 3 HIGH: 0					
Great Place To Work	Last revised: 2021-04-03 09:42:32		Training examples: 10		
LOW: 44 MED: 41 HIGH: 111					
Stress	Last revised: 2021-04-05 06:05:14		Training examples: 167		
LOW: 75 MED: 27 HIGH: 30					
Customer Focus	Last revised: 2021-04-03 09:42:32		Training examples: 9		
LOW: 109 MED: 75 HIGH: 185					
Alignment	Last revised: 2021-04-03 09:42:32		Training examples: 0		
LOW: 0 MED: 1 HIGH: 4					
Company Car	Last revised: 2021-04-05 06:16:22		Training examples: 15		
LOW: 4 MED: 1 HIGH: 0					

Andrew Marritt | Founder & CEO

andrew.marritt@organizationview.com | +41 78 905 0775