TASK 2: MODEL TRAINING

Ching-Han KUO

PREPROCESSING

STEPS

- Make all characters in the lower cases
- Tokenise the text (nltk.word_tokenize)
- Lemmatize the text (WordNetLemmatizer)
- Eliminate stop words (ENGLISH_STOP_WORDS)
- Eliminate punctuation (string)

EXAMPLE

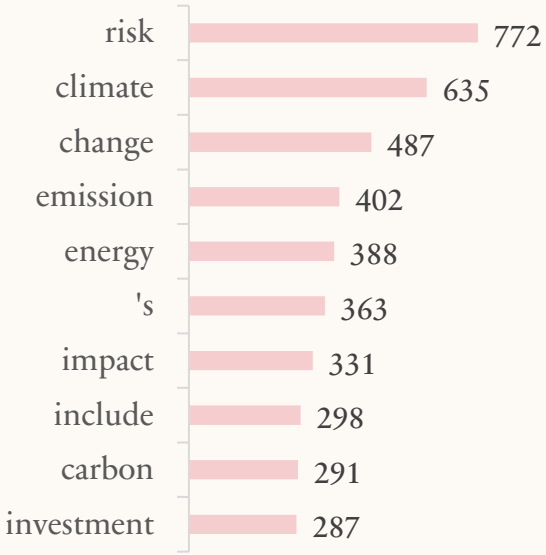
- – Scope 3: Optional scope that includes indirect emissions associated with the goods and services supply chain produced outside the organization.
- scope 3 optional scope include indirect emission associate good service supply chain produce outside organization

	text	label	processed	processed_wo_punct
0	– Scope 3: Optional scope that includes indire...	1	[–, scope, 3, :, optional, scope, include, ind...	[scope, 3, optional, scope, include, indirect,...
1	The Group is not aware of any noise pollution ...	0	[group, aware, noise, pollution, negatively, i...	[group, aware, noise, pollution, negatively, i...
2	Global climate change could exacerbate certain...	0	[global, climate, change, exacerbate, certain,...	[global, climate, change, exacerbate, certain,...
3	Setting an investment horizon is part and parc...	0	[set, investment, horizon, parcel, policy, foc...	[set, investment, horizon, parcel, policy, foc...
4	Climate change the physical impacts of climate...	0	[climate, change, physical, impact, climate, c...	[climate, change, physical, impact, climate, c...

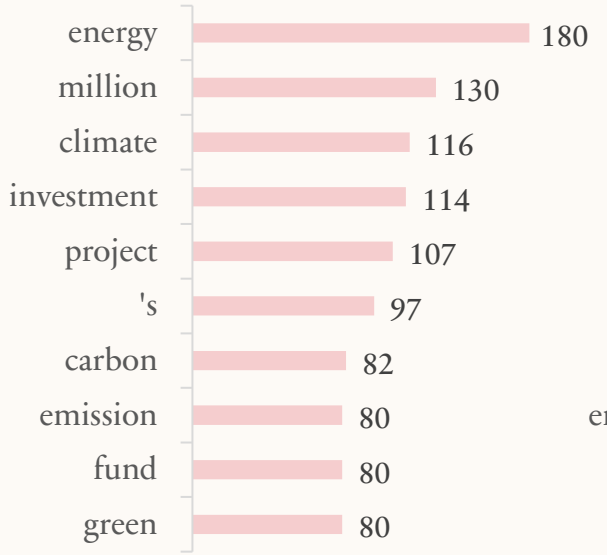
DESCRIPTIVE STATISTICS

Word Frequency: Top 10

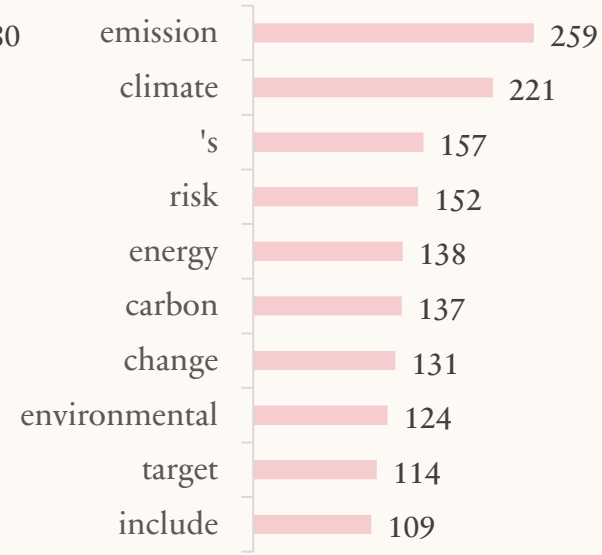
All



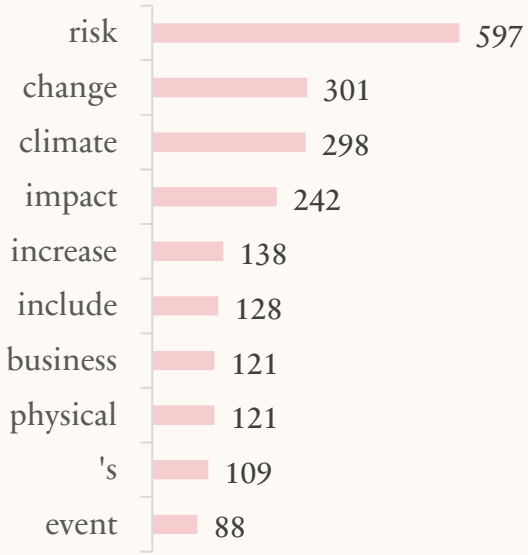
Positive



Neutral

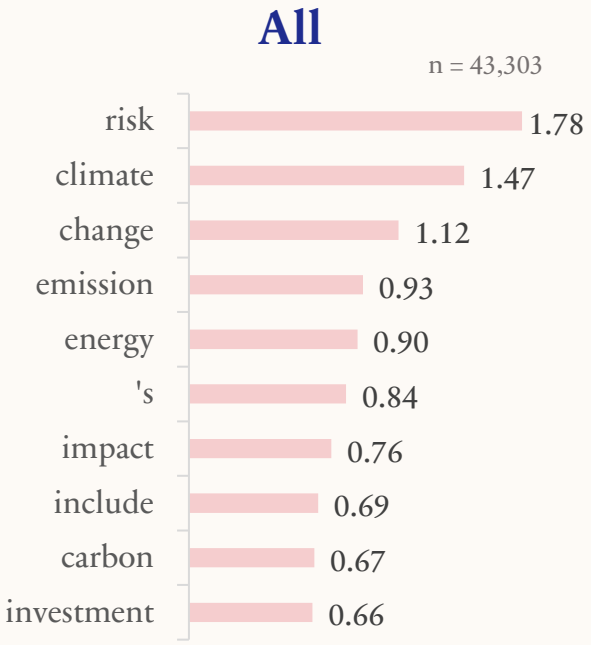


Negative



DESCRIPTIVE STATISTICS

Word Frequency: z-test



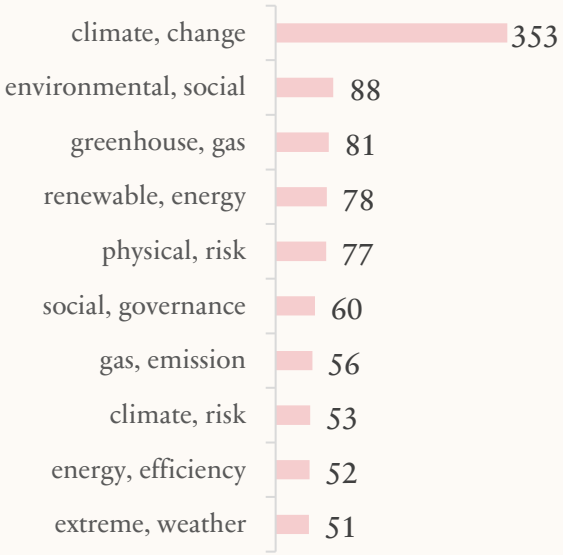
	Positive	Neutral	Negative
n=	*11,597*	*16,730*	*16,976*
	0.20	0.91	3.99
	1.00	1.32	1.99
	0.47	0.78	2.01
	0.69	1.55	0.42
	1.55	0.82	0.47
	0.84	0.94	0.73
	0.22	0.38	1.62
	0.53	0.65	0.85
	0.71	0.82	0.48
	0.98	0.63	0.45

* Red/Green indicates it is significantly higher/lower than other groups (proportional z-test)

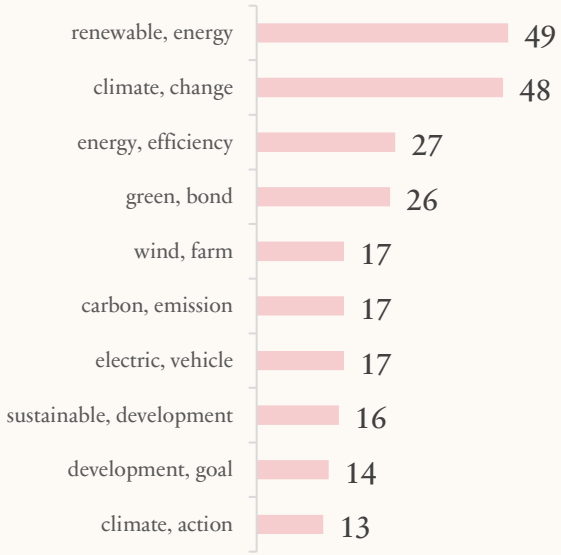
DESCRIPTIVE STATISTICS

Bi-gram Frequency: Top 10

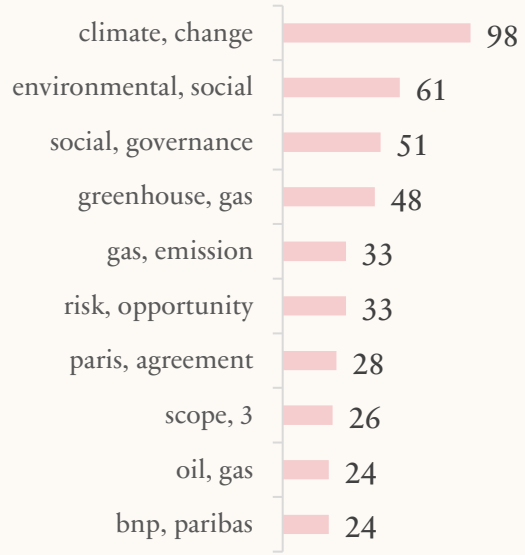
All



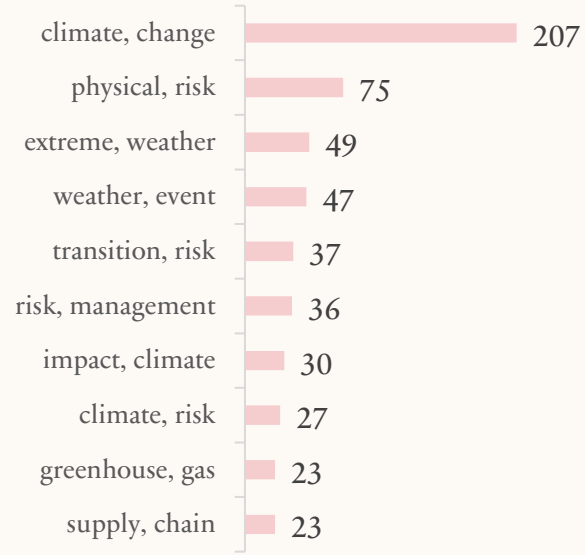
Positive



Neutral



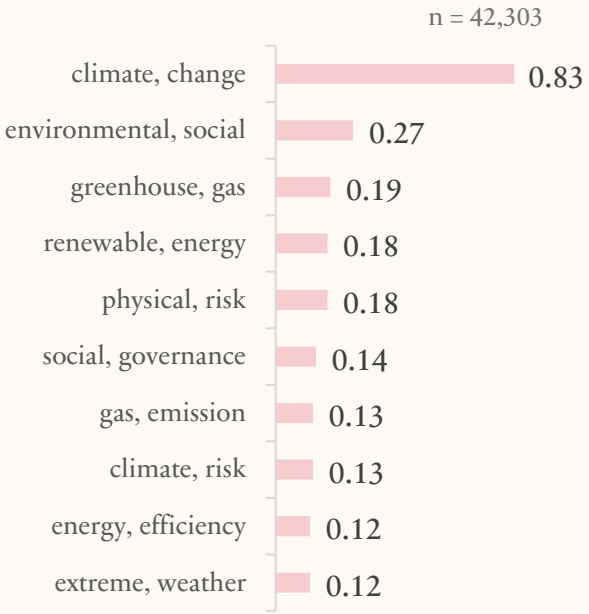
Negative



DESCRIPTIVE STATISTICS

Bi-gram Frequency: z-test

All

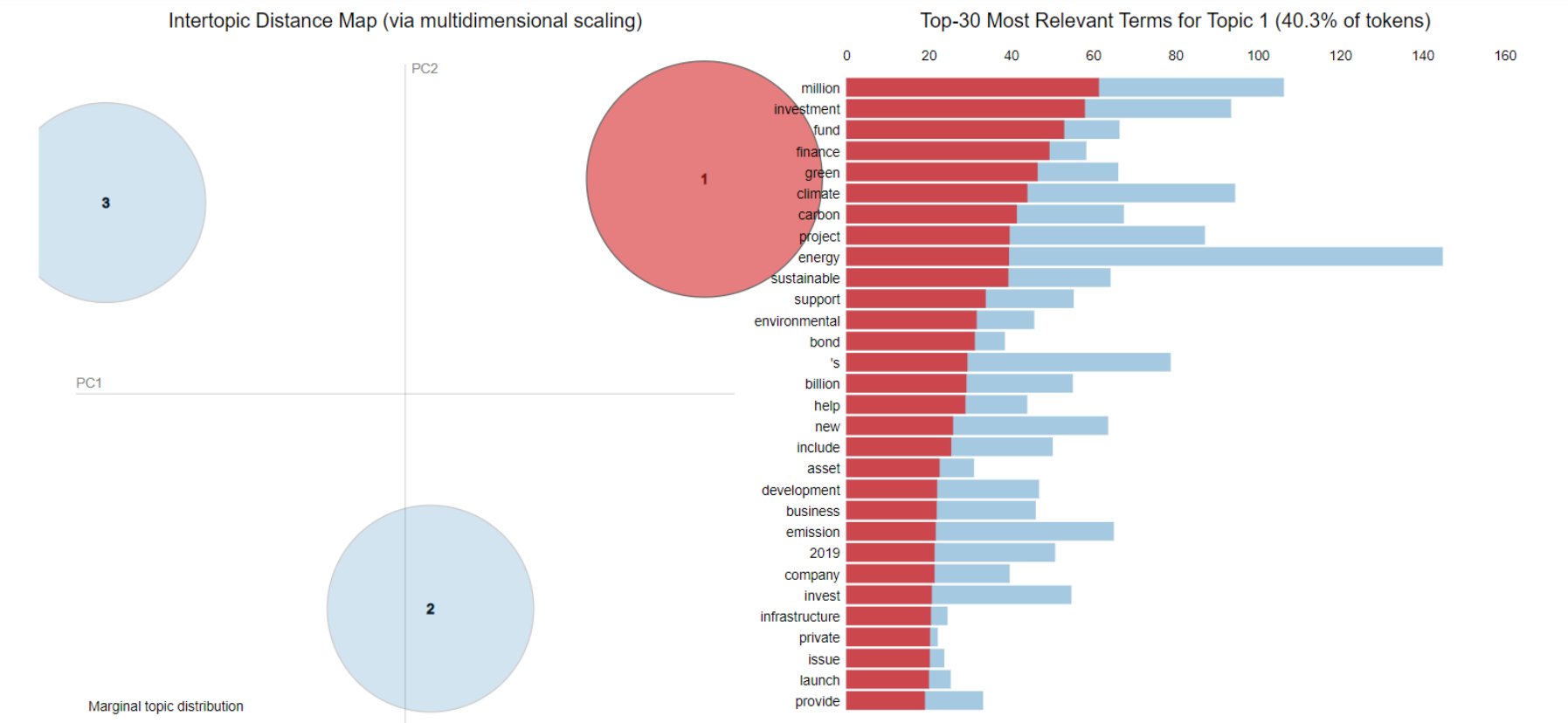


	Positive	Neutral	Negative
n=	*11,347*	*16,322*	*14,634*
	0.42	0.60	1.41
	0.09	0.37	0.12
	0.09	0.29	0.16
	0.43	0.12	0.06
	0.00	0.01	0.51
	0.03	0.31	0.04
	0.07	0.20	0.10
	0.04	0.13	0.18
	0.24	0.13	0.02
	0.00	0.01	0.33

* Red/Green indicates it is significantly higher/lower than other groups (proportional z-test)

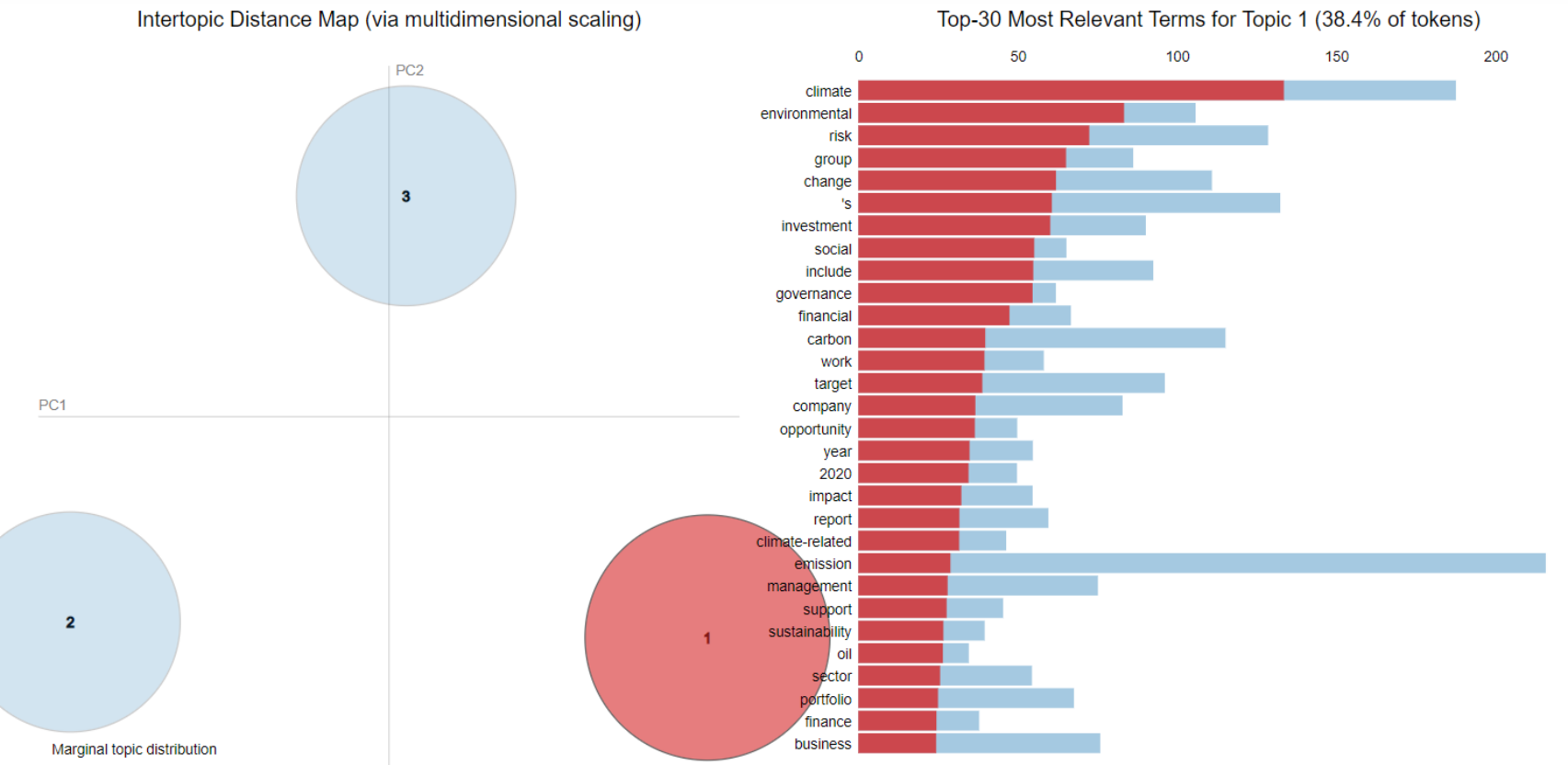
DESCRIPTIVE STATISTICS

Topic Modelling: Positive



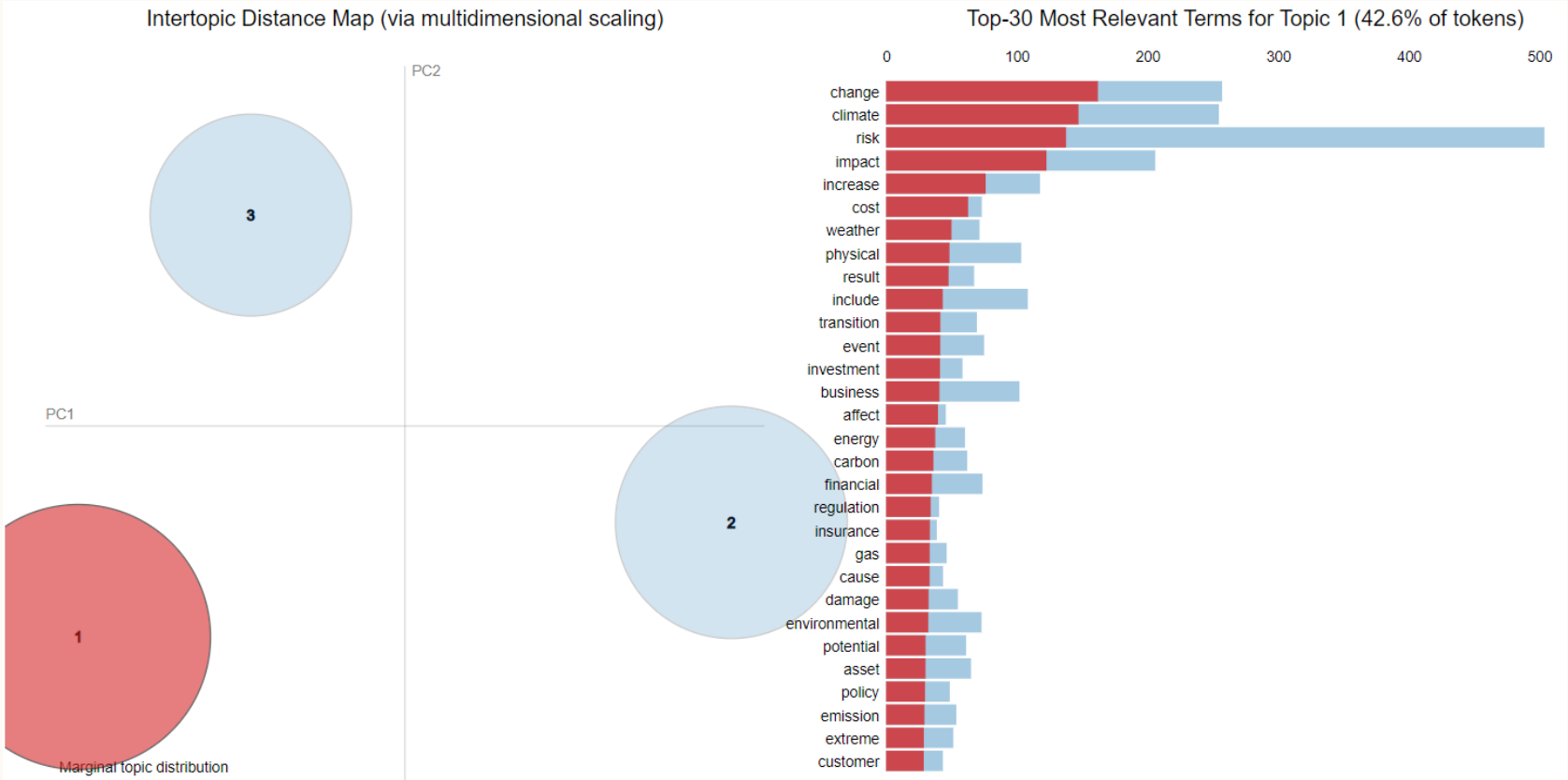
DESCRIPTIVE STATISTICS

Topic Modelling: Neutral



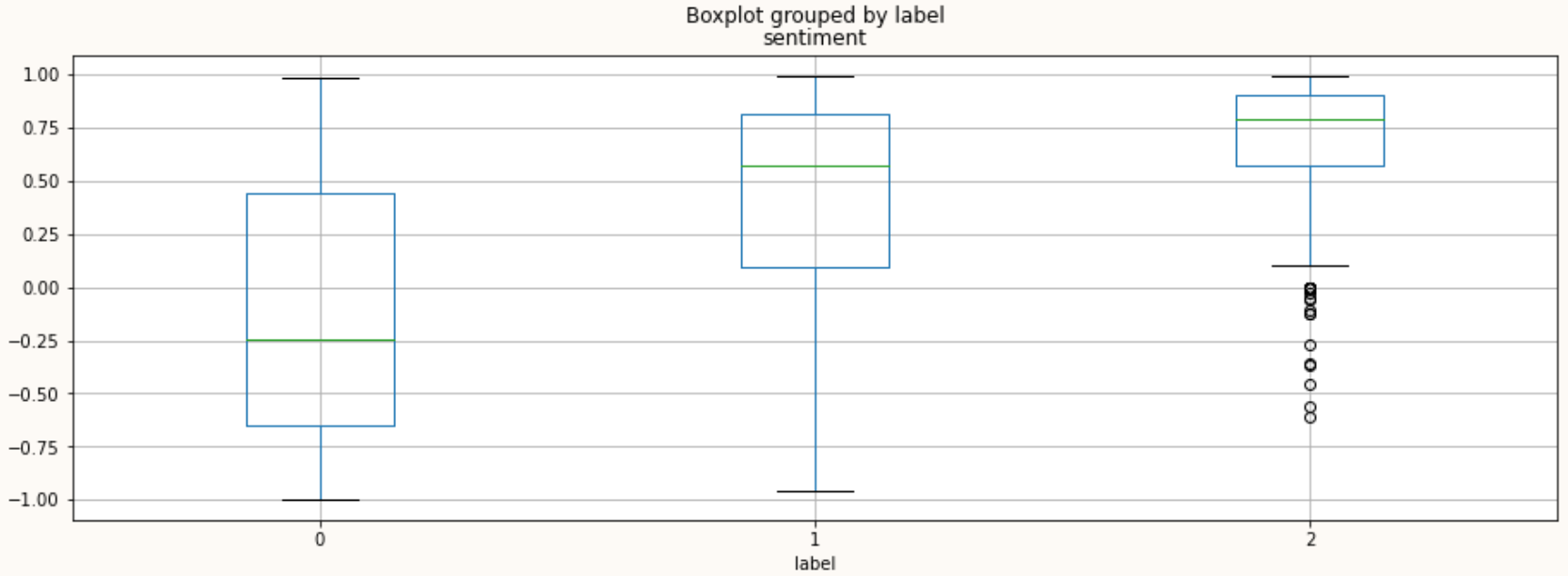
DESCRIPTIVE STATISTICS

Topic Modelling: Negative



DESCRIPTIVE STATISTICS

Sentiment Analysis



correlation: 0.5401455003938614, p-value: 8.121214139816934e-77

PREDICTION MODEL

FEATURE

- Bag of Words
- Doc2Vec (Word2Vec)
- Sentiment Score

MODEL TYPE

- Baseline: Dummy (most frequent)
- Logistic Regression
- Decision Tree

	bow_000	bow_000m	bow_000m3	bow_000t	bow_01	bow_057	bow_07	bow_08	bow_088	bow_09	...	w2v_94	w2v_95	w2v_96	w2v_97
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.253211	0.083096	0.078186	-0.100423
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.382230	0.133701	0.103212	-0.134507
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.549992	0.193376	0.148590	-0.216368
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.471259	0.152975	0.134510	-0.203176
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.252683	0.088680	0.064720	-0.090843
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
315	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.159054	0.049949	0.043138	-0.079112
316	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.321937	0.124167	0.102547	-0.119688
317	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.325235	0.113128	0.094136	-0.131708
318	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.315236	0.114299	0.092415	-0.119096
319	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.361904	0.192984	0.161625	-0.140768

MODEL EVALUATION

	Dummy		Logistic Regression		Decision Tree	
	Score	95% CI	Score	95% CI	Score	95% CI
10 CV	0.41	0.03	0.73	0.03	0.66	0.03
Accuracy	0.50	0.05	0.76	0.05	0.61	0.05
Precision	0.17	0.04	0.72	0.05	0.57	0.05
Recall	0.33	0.05	0.73	0.05	0.59	0.05
F1-score	0.22	0.04	0.73	0.05	0.58	0.05

THANK YOU

Ching-Han KUO
kuochinhan@gmail.com
ching-han.kuo@student.kuleuven.be