

# Web content sentimental analysis: News

**Karolina Bzdusek**

Capstone Project for Intensive Data Science Career Track, July 22nd 2019

# Introduction

Source of data: <https://rishabhmisra.github.io/publications/>

	authors	category	date	headline	link	short_description
0	Melissa Jeltsen	CRIME	2018-05-26	There Were 2 Mass Shootings In Texas Last Week...	<a href="https://www.huffingtonpost.com/entry/texas-ama...">https://www.huffingtonpost.com/entry/texas-ama...</a>	She left her husband. He killed their children...
1	Andy McDonald	ENTERTAINMENT	2018-05-26	Will Smith Joins Diplo And Nicky Jam For The 2...	<a href="https://www.huffingtonpost.com/entry/will-smit...">https://www.huffingtonpost.com/entry/will-smit...</a>	Of course it has a song.
2	Ron Dicker	ENTERTAINMENT	2018-05-26	Hugh Grant Marries For The First Time At Age 57	<a href="https://www.huffingtonpost.com/entry/hugh-gran...">https://www.huffingtonpost.com/entry/hugh-gran...</a>	The actor and his longtime girlfriend Anna Ebe...
3	Ron Dicker	ENTERTAINMENT	2018-05-26	Jim Carrey Blasts 'Castrato' Adam Schiff And D...	<a href="https://www.huffingtonpost.com/entry/jim-carre...">https://www.huffingtonpost.com/entry/jim-carre...</a>	The actor gives Dems an ass-kicking for not fi...
4	Ron Dicker	ENTERTAINMENT	2018-05-26	Julianna Margulies Uses Donald Trump Poop Bags...	<a href="https://www.huffingtonpost.com/entry/julianna-...">https://www.huffingtonpost.com/entry/julianna-...</a>	The "Dietland" actress said using the bags is ...
5	Ron Dicker	ENTERTAINMENT	2018-05-26	Morgan Freeman 'Devastated' That Sexual Harass...	<a href="https://www.huffingtonpost.com/entry/morgan-fr...">https://www.huffingtonpost.com/entry/morgan-fr...</a>	"It is not right to equate horrific incidents ...
6	Ron Dicker	ENTERTAINMENT	2018-05-26	Donald Trump Is Lovin' New McDonald's Jingle I...	<a href="https://www.huffingtonpost.com/entry/donald-tr...">https://www.huffingtonpost.com/entry/donald-tr...</a>	It's catchy, all right.
7	Todd Van Luling	ENTERTAINMENT	2018-05-26	What To Watch On Amazon Prime That's New This ...	<a href="https://www.huffingtonpost.com/entry/amazon-pr...">https://www.huffingtonpost.com/entry/amazon-pr...</a>	There's a great mini-series joining this week.
8	Andy McDonald	ENTERTAINMENT	2018-05-26	Mike Myers Reveals He'd 'Like To' Do A Fourth ...	<a href="https://www.huffingtonpost.com/entry/mike-myer...">https://www.huffingtonpost.com/entry/mike-myer...</a>	Myer's kids may be pushing for a new "Powers" ...
9	Todd Van Luling	ENTERTAINMENT	2018-05-26	What To Watch On Hulu That's New This Week	<a href="https://www.huffingtonpost.com/entry/hulu-what...">https://www.huffingtonpost.com/entry/hulu-what...</a>	You're getting a recent Academy Award-winning ...

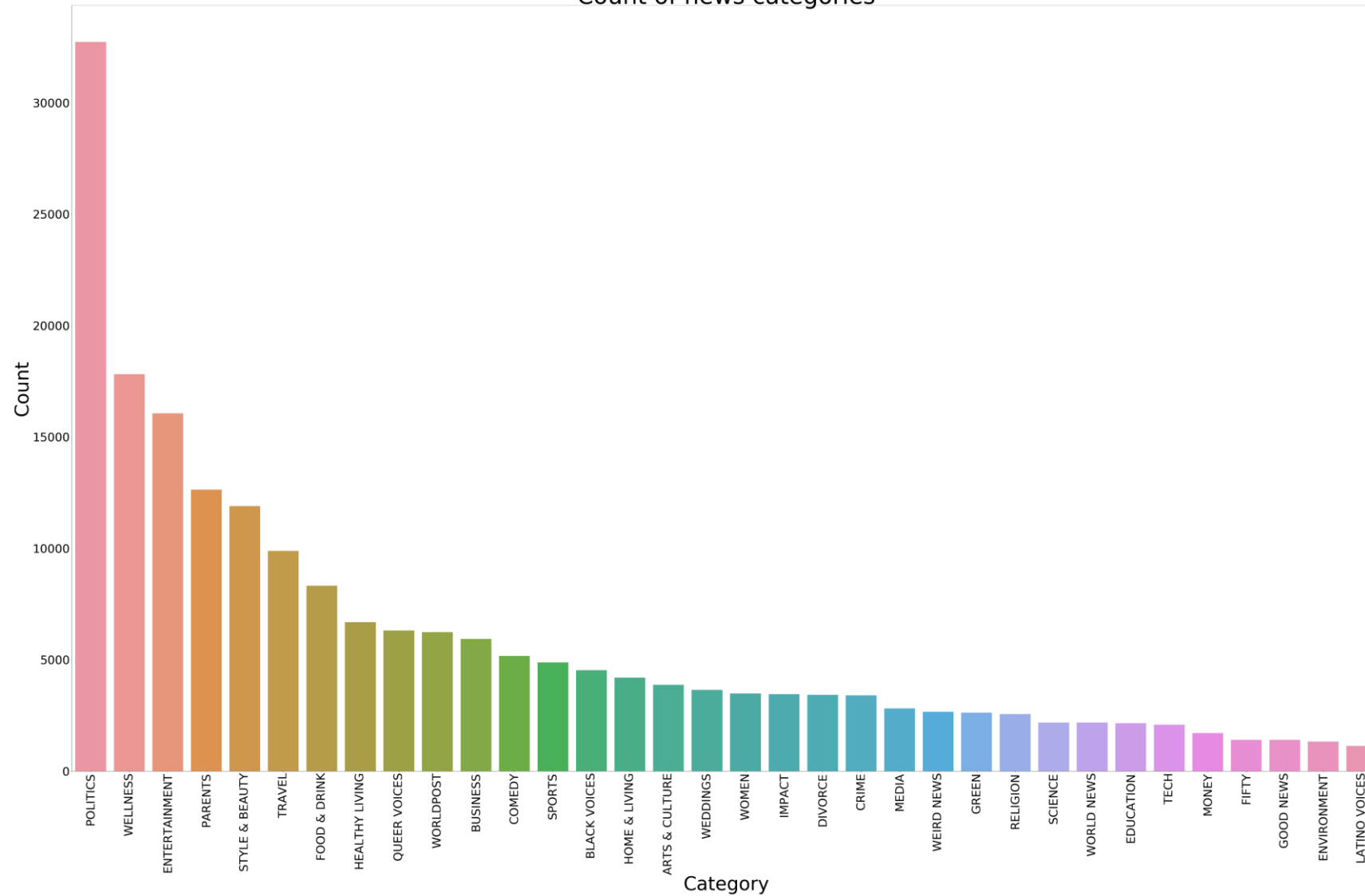
# Introduction

Source of data: <https://rishabhmisra.github.io/publications/>

	authors	category	date	headline	link	short_description
0	Melissa Jeltsen	CRIME	2018-05-26	There Were 2 Mass Shootings In Texas Last Week...	<a href="https://www.huffingtonpost.com/entry/texas-ama...">https://www.huffingtonpost.com/entry/texas-ama...</a>	She left her husband. He killed their children...
1	Andy McDonald	ENTERTAINMENT	2018-05-26	Will Smith Joins Diplo And Nicky Jam For The 2...	<a href="https://www.huffingtonpost.com/entry/will-smit...">https://www.huffingtonpost.com/entry/will-smit...</a>	Of course it has a song.
2	Ron Dicker	ENTERTAINMENT	2018-05-26	Hugh Grant Marries For The First Time At Age 57	<a href="https://www.huffingtonpost.com/entry/hugh-gran...">https://www.huffingtonpost.com/entry/hugh-gran...</a>	The actor and his longtime girlfriend Anna Ebe...
3	Ron Dicker	ENTERTAINMENT	2018-05-26	Jim Carrey Blasts 'Castrato' Adam Schiff And D...	<a href="https://www.huffingtonpost.com/entry/jim-carre...">https://www.huffingtonpost.com/entry/jim-carre...</a>	The actor gives Dems an ass-kicking for not fi...
4	Ron Dicker	ENTERTAINMENT	2018-05-26	Julianna Margulies Uses Donald Trump Poop Bags...	<a href="https://www.huffingtonpost.com/entry/julianna-...">https://www.huffingtonpost.com/entry/julianna-...</a>	The "Dietland" actress said using the bags is ...
5	Ron Dicker	ENTERTAINMENT	2018-05-26	Morgan Freeman 'Devastated' That Sexual Harass...	<a href="https://www.huffingtonpost.com/entry/morgan-fr...">https://www.huffingtonpost.com/entry/morgan-fr...</a>	"It is not right to equate horrific incidents ...
6	Ron Dicker	ENTERTAINMENT	2018-05-26	Donald Trump Is Lovin' New McDonald's Jingle I...	<a href="https://www.huffingtonpost.com/entry/donald-tr...">https://www.huffingtonpost.com/entry/donald-tr...</a>	It's catchy, all right.
7	Todd Van Luling	ENTERTAINMENT	2018-05-26	What To Watch On Amazon Prime That's New This ...	<a href="https://www.huffingtonpost.com/entry/amazon-pr...">https://www.huffingtonpost.com/entry/amazon-pr...</a>	There's a great mini-series joining this week.
8	Andy McDonald	ENTERTAINMENT	2018-05-26	Mike Myers Reveals He'd 'Like To' Do A Fourth ...	<a href="https://www.huffingtonpost.com/entry/mike-myer...">https://www.huffingtonpost.com/entry/mike-myer...</a>	Myer's kids may be pushing for a new "Powers" ...
9	Todd Van Luling	ENTERTAINMENT	2018-05-26	What To Watch On Hulu That's New This Week	<a href="https://www.huffingtonpost.com/entry/hulu-what...">https://www.huffingtonpost.com/entry/hulu-what...</a>	You're getting a recent Academy Award-winning ...

# Introduction

Count of news categories



# Introduction

41 categories → 34 categories

- 'THE WORLDPOST', 'WORLDPOST'
- 'ARTS', 'CULTURE & ARTS', 'ARTS & CULTURE'
- 'PARENTING', 'PARENTS'
- 'STYLE', 'STYLE & BEAUTY'
- 'COLLEGE', 'EDUCATION'
- 'TASTE', 'FOOD & DRINK'

Category "POLITICS"

Data\_pol

	total_count
count	20922.0
mean	17.48891119395851
std	162.74761971155192
min	1.0
25%	1.0
50%	2.0
75%	6.0
max	9180.0

# Models

1. TFIDF matrix
2. Document vectors matrix

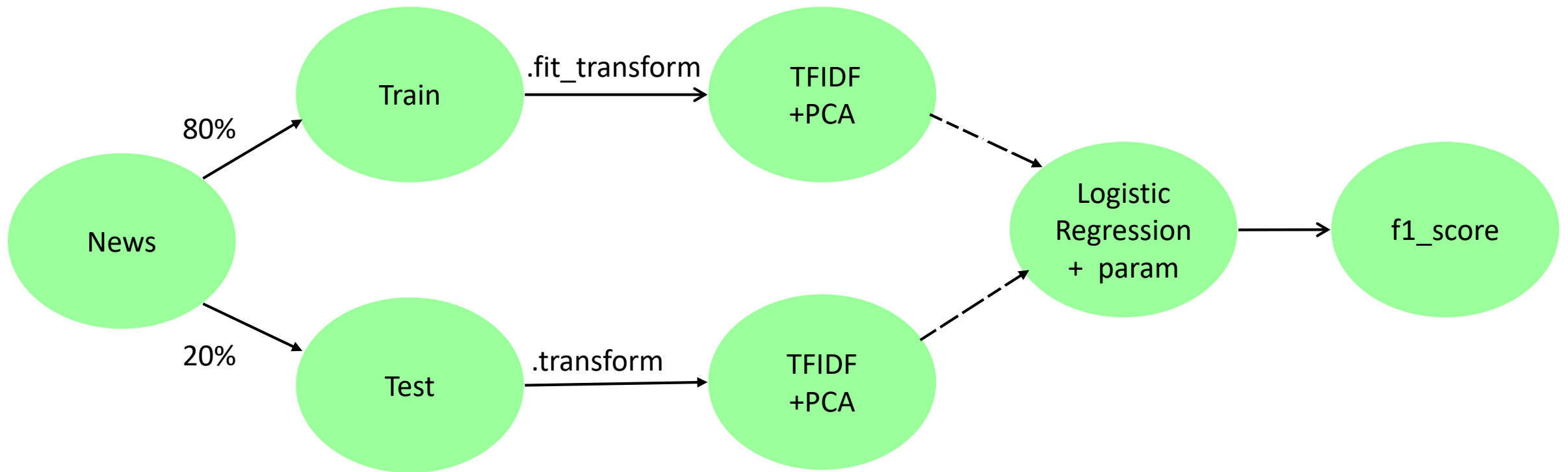
## TFIDF hyperparameters

- NLP feature selection
- TweetTokenizer
- PCA
- Solver, max\_df, ngram range

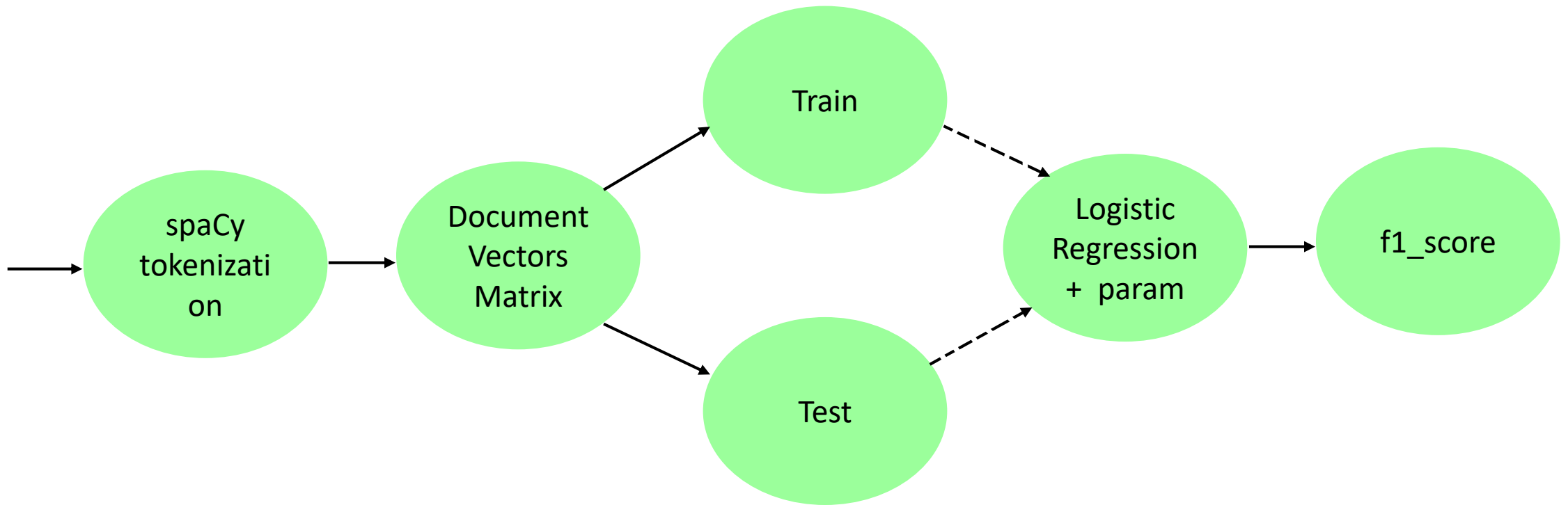
## Document vectors

- spaCy tokenization
- 'en\_core\_web\_lg'

# Modelling TFIDF matrix

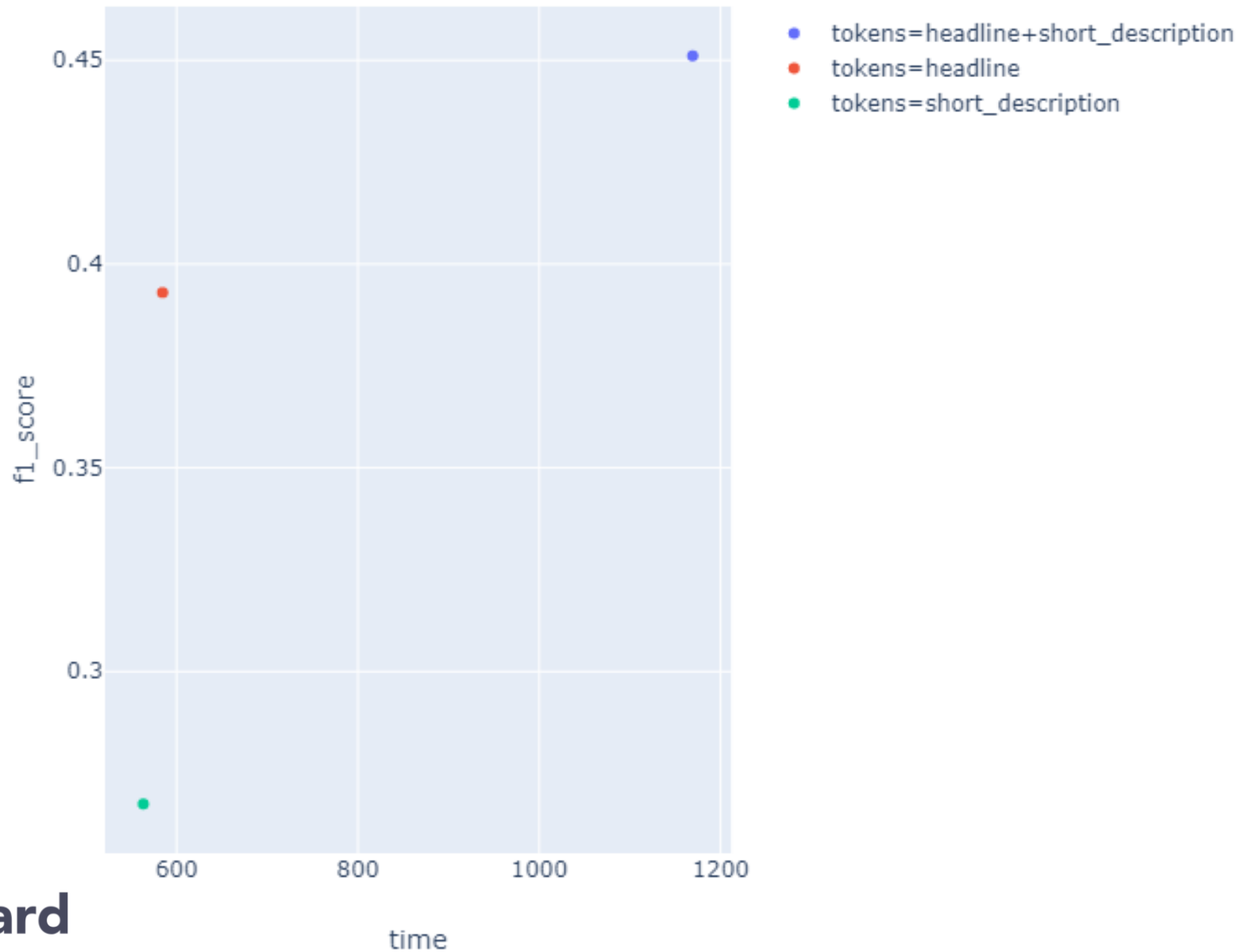


# Modelling Document Vectors

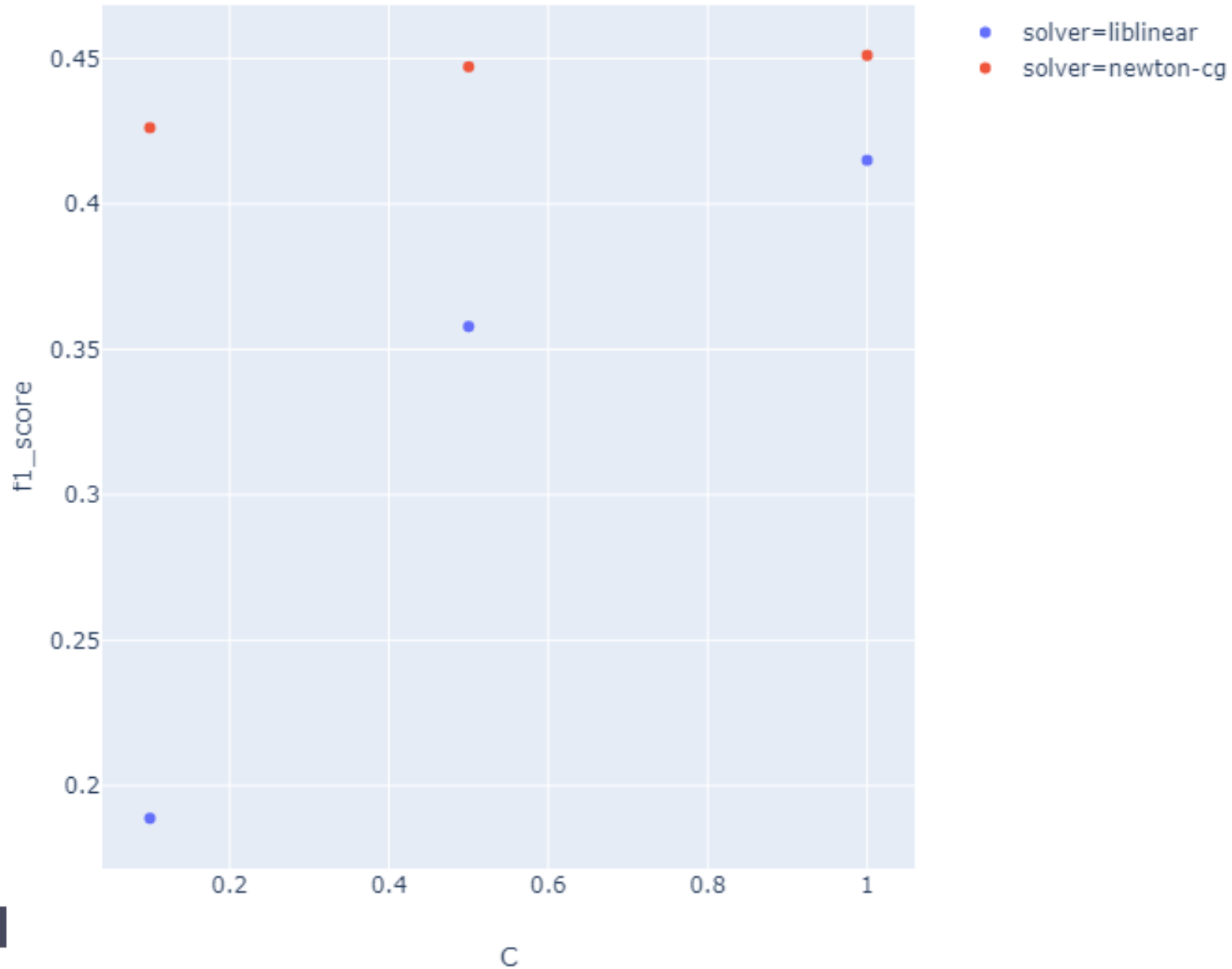




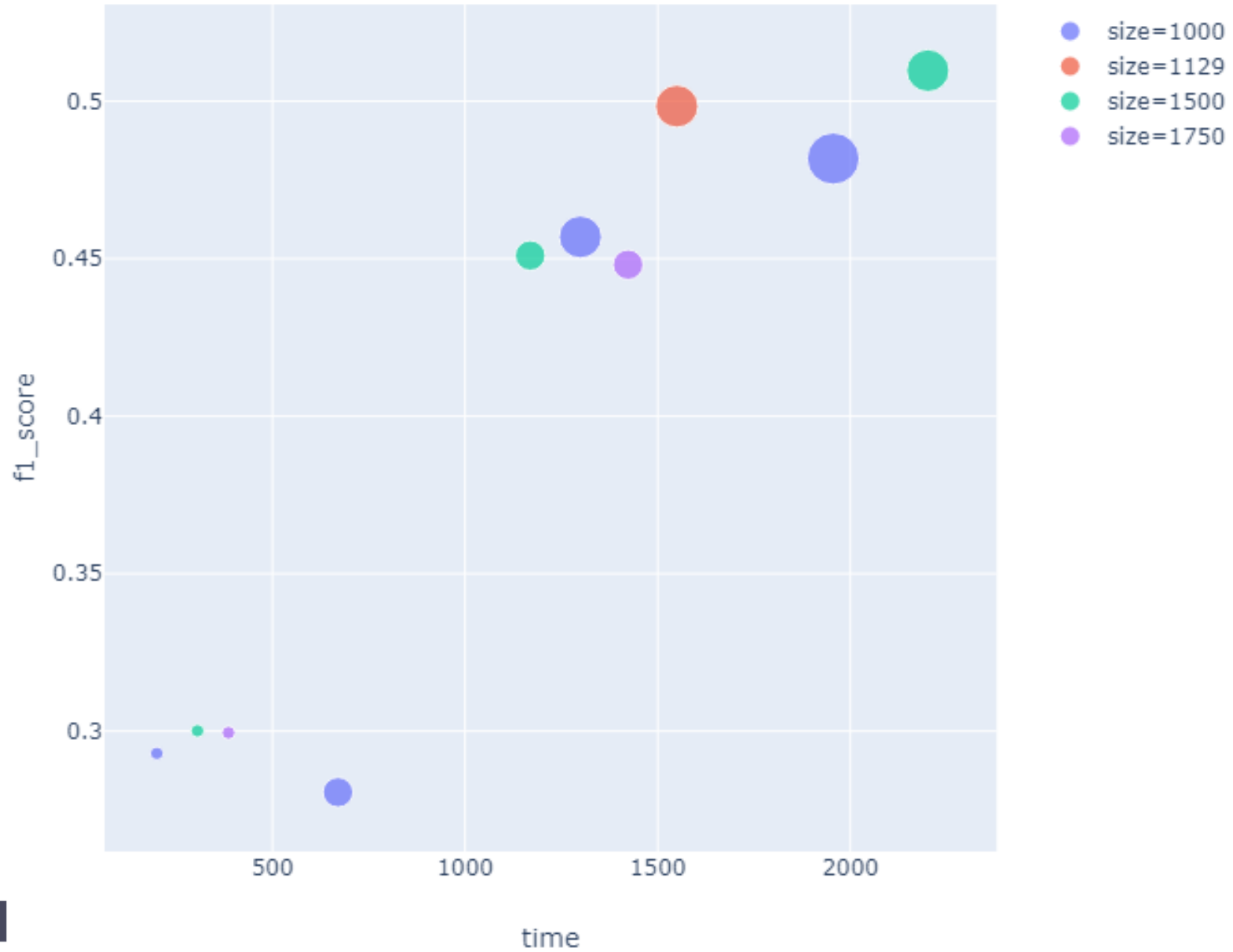
# Results



# Results



# Results



# Results



# Results

	F1_score	Runtime
TFIDF	0.4306	309.1294
Document matrix	0.4676	490.0666

*Our computation is performed on a subset of the original dataset (first 40,000 rows). Therefore we are predicting not 34 categories, just 27.*

# Conclusion

- Model Comparison
- Future work
  - Sample size (working in batches to manage oversampling)
  - Use neural networks with word vectors
  - Use different word vectors

# Thank you!

(Q&A)