# Web content sentiment analysis: News

**Karolina Bzdusek**

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

Springboard

# 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... | https://www.huffingtonpost.com/entry/texas-ama... | 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... | https://www.huffingtonpost.com/entry/will-smit... | Of course it has a song. |
| **2** | Ron Dicker | ENTERTAINMENT | 2018-05-26 | Hugh Grant Marries For The First Time At Age 57 | https://www.huffingtonpost.com/entry/hugh-gran... | The actor and his longtime girlfriend Anna Ebe... |
| **3** | Ron Dicker | ENTERTAINMENT | 2018-05-26 | Jim Carrey Blasts 'Castrato' Adam Schiff And D... | https://www.huffingtonpost.com/entry/jim-carre... | The actor gives Dems an ass-kicking for not fi... |
| **4** | Ron Dicker | ENTERTAINMENT | 2018-05-26 | Julianna Margulies Uses Donald Trump Poop Bags... | https://www.huffingtonpost.com/entry/julianna-... | The "Dietland" actress said using the bags is ... |
| **5** | Ron Dicker | ENTERTAINMENT | 2018-05-26 | Morgan Freeman 'Devastated' That Sexual Harass... | https://www.huffingtonpost.com/entry/morgan-fr... | "It is not right to equate horrific incidents ... |
| **6** | Ron Dicker | ENTERTAINMENT | 2018-05-26 | Donald Trump Is Lovin' New McDonald's Jingle I... | https://www.huffingtonpost.com/entry/donald-tr... | It's catchy, all right. |
| **7** | Todd Van Luling | ENTERTAINMENT | 2018-05-26 | What To Watch On Amazon Prime That's New This ... | https://www.huffingtonpost.com/entry/amazon-pr... | 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 ... | https://www.huffingtonpost.com/entry/mike-myer... | 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 | https://www.huffingtonpost.com/entry/hulu-what... | You're getting a recent Academy Award-winning ... |



Karolina Bzdusek,
22nd July 2019

# Introduction

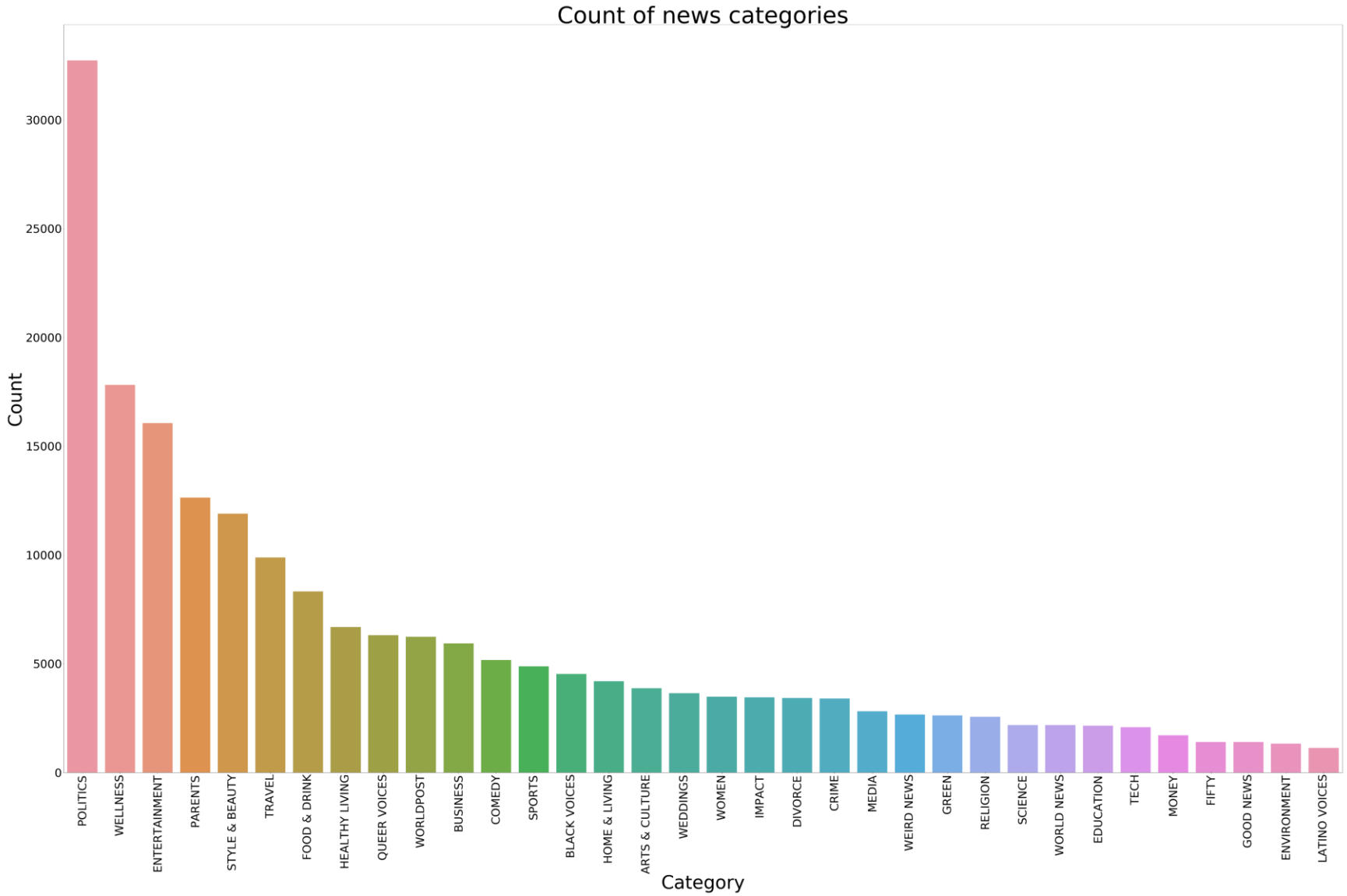
Source of data: https://rishabhmisra.github.io/publications/

| | authors | category | date | headline | link | short_description |
|---|---|---|---|---|---|---|
| 0 | Melissa Jeltsen | CRIME | 2018-05-26 | There Were 2 Mass Shootings In Texas Last Week... | https://www.huffingtonpost.com/entry/texas-ama... | 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... | https://www.huffingtonpost.com/entry/will-smit... | Of course it has a song. |
| 2 | Ron Dicker | ENTERTAINMENT | 2018-05-26 | Hugh Grant Marries For The First Time At Age 57 | https://www.huffingtonpost.com/entry/hugh-gran... | The actor and his longtime girlfriend Anna Ebe... |
| 3 | Ron Dicker | ENTERTAINMENT | 2018-05-26 | Jim Carrey Blasts 'Castrato' Adam Schiff And D... | https://www.huffingtonpost.com/entry/jim-carre... | The actor gives Dems an ass-kicking for not fi... |
| 4 | Ron Dicker | ENTERTAINMENT | 2018-05-26 | Julianna Margulies Uses Donald Trump Poop Bags... | https://www.huffingtonpost.com/entry/julianna-... | The "Dietland" actress said using the bags is ... |
| 5 | Ron Dicker | ENTERTAINMENT | 2018-05-26 | Morgan Freeman 'Devastated' That Sexual Harass... | https://www.huffingtonpost.com/entry/morgan-fr... | "It is not right to equate horrific incidents ... |
| 6 | Ron Dicker | ENTERTAINMENT | 2018-05-26 | Donald Trump Is Lovin' New McDonald's Jingle I... | https://www.huffingtonpost.com/entry/donald-tr... | It's catchy, all right. |
| 7 | Todd Van Luling | ENTERTAINMENT | 2018-05-26 | What To Watch On Amazon Prime That's New This ... | https://www.huffingtonpost.com/entry/amazon-pr... | 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 ... | https://www.huffingtonpost.com/entry/mike-myer... | 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 | https://www.huffingtonpost.com/entry/hulu-what... | You're getting a recent Academy Award-winning ... |

Springboard

Karolina Bzdusek,
22nd July 2019

# Introduction




Springboard

Karolina Bzdusek,
22nd July 2019

# 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 | 17.488911119395851 |
| std | 162.74761971155192 |
| min | 1.0 |
| 25% | 1.0 |
| 50% | 2.0 |
| 75% | 6.0 |
| max | 9180.0 |

Springboard

Karolina Bzdusek,
22nd July 2019

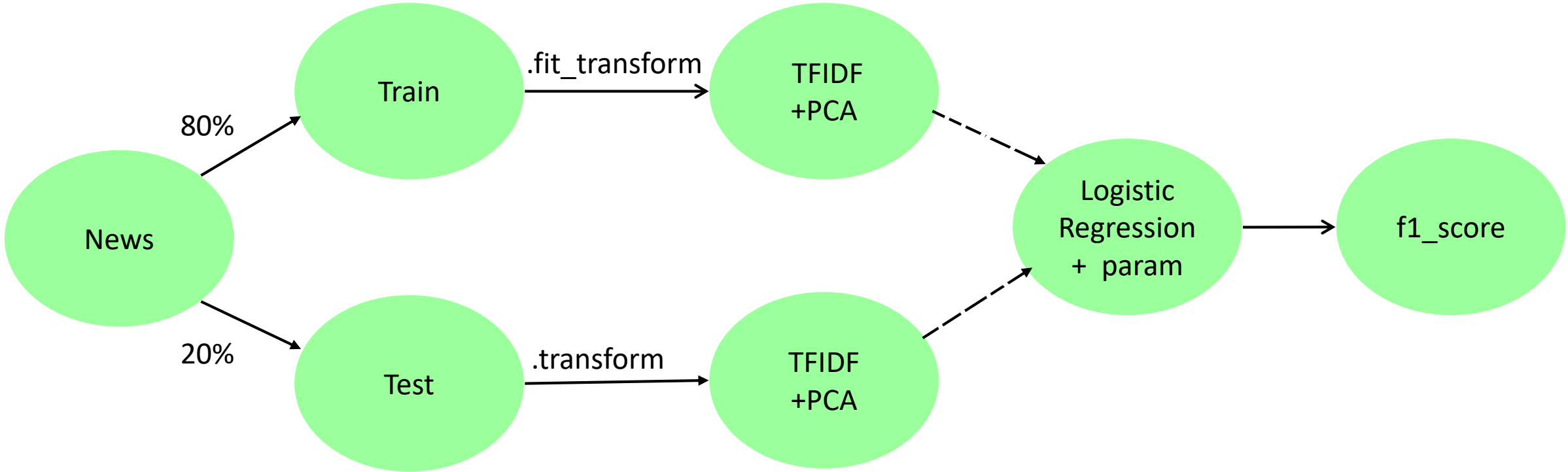
# Models

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’

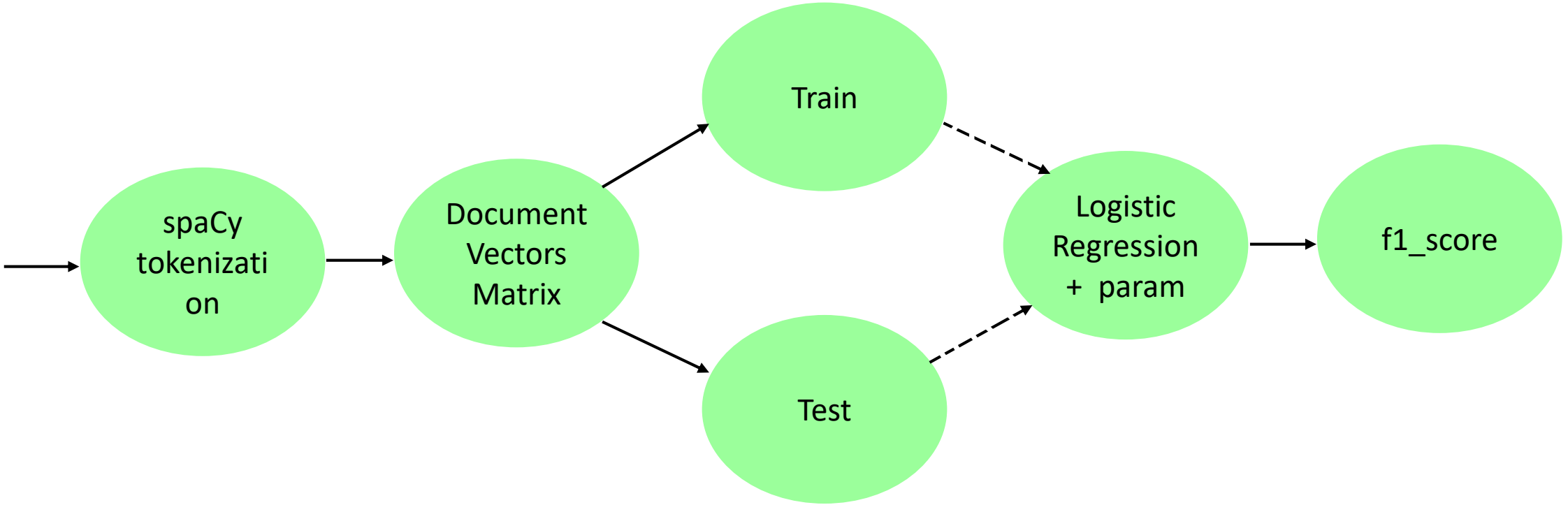
# Modelling TFIDF matrix




Springboard

Karolina Bzdusek,
20th May 2019

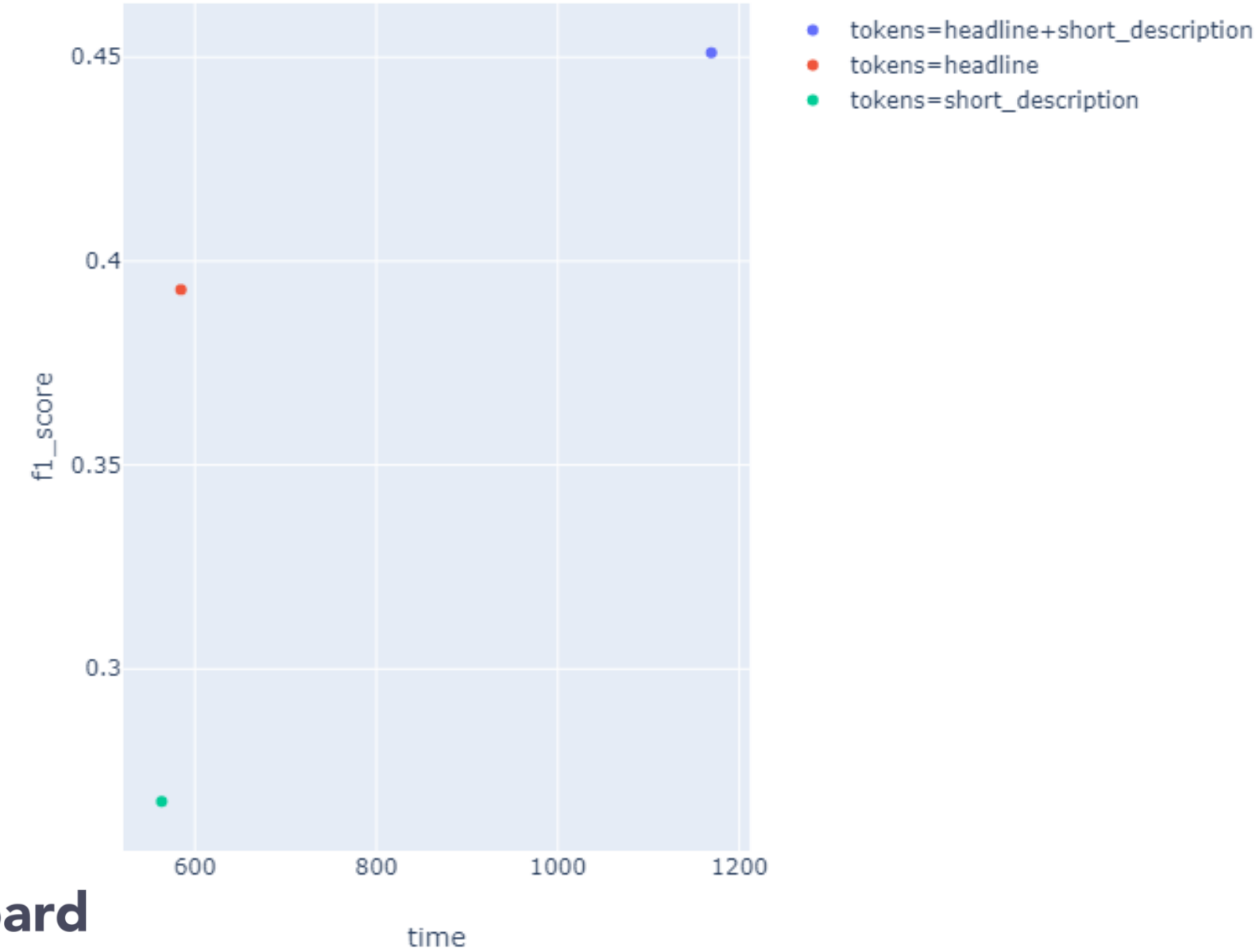
# Modelling Document Vectors




Springboard

Karolina Bzdusek,
20th May 2019

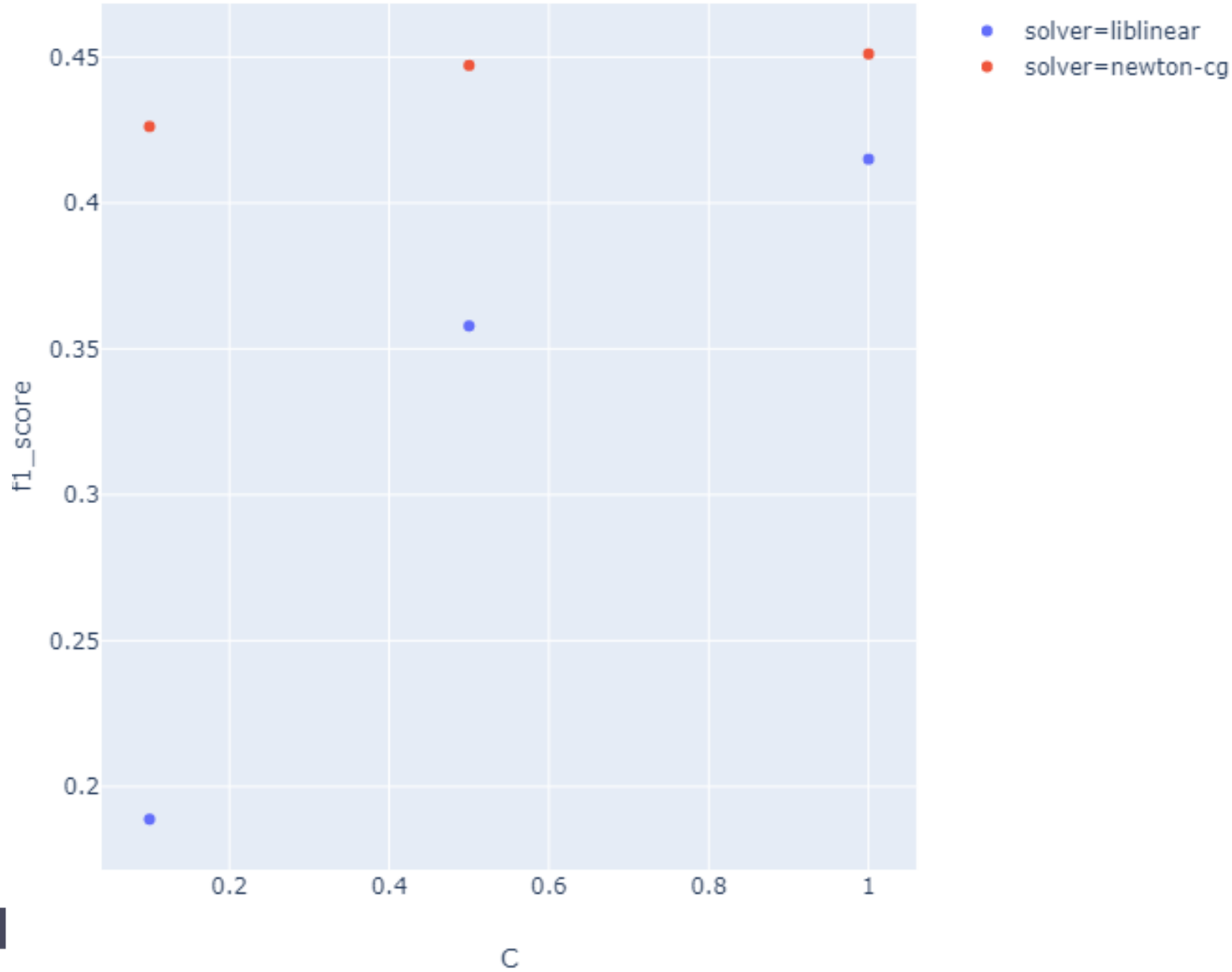
# Results




Springboard

Karolina Bzdusek,
22nd July 2019

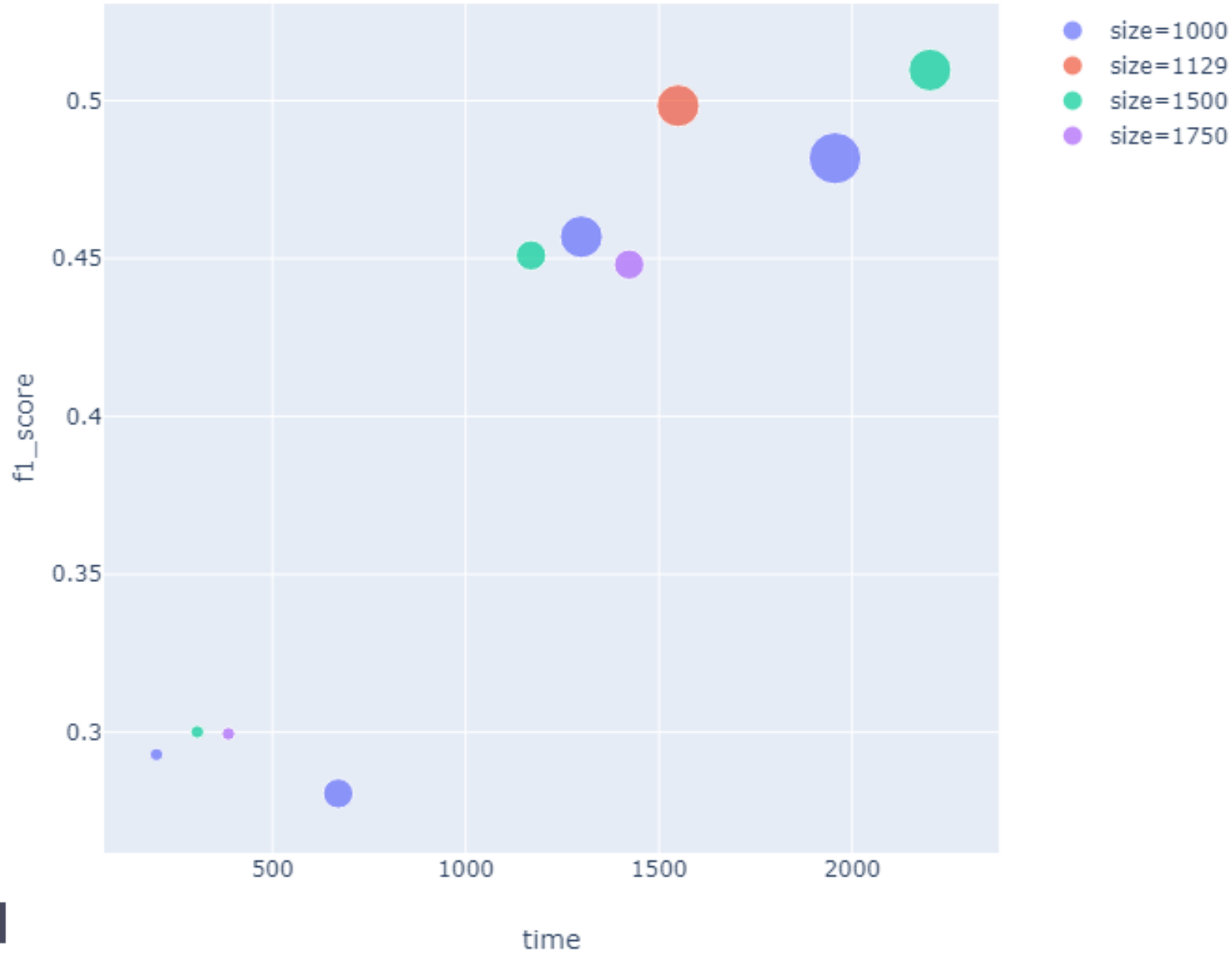
# Results




Springboard

Karolina Bzdusek,
22nd July 2019

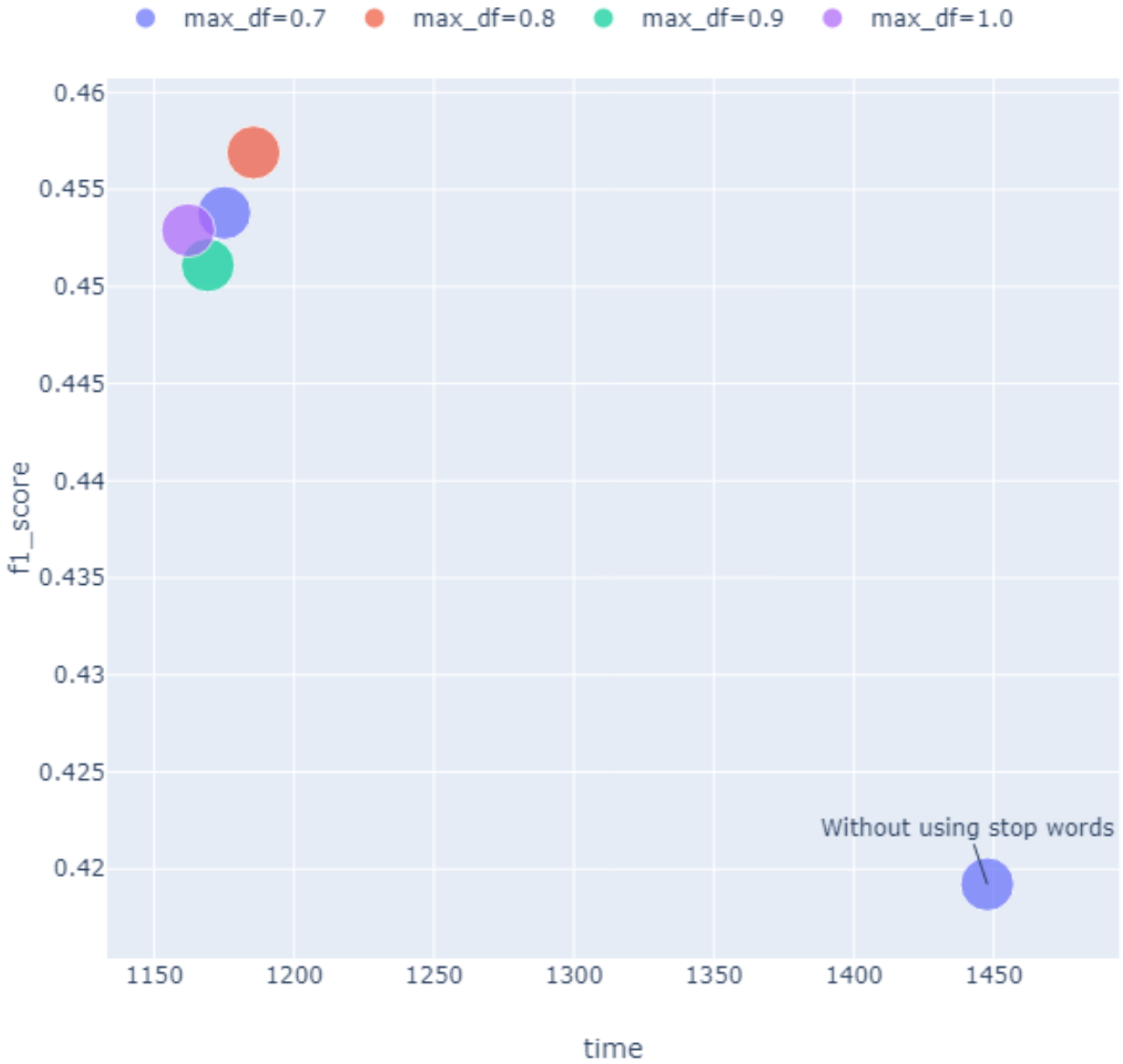
# Results




Springboard

Karolina Bzdusek,
22nd July 2019

# Results




Springboard

Karolina Bzdusek,
22nd July 2019

# 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)

Springboard

Karolina Bzdusek,
20th May 2019