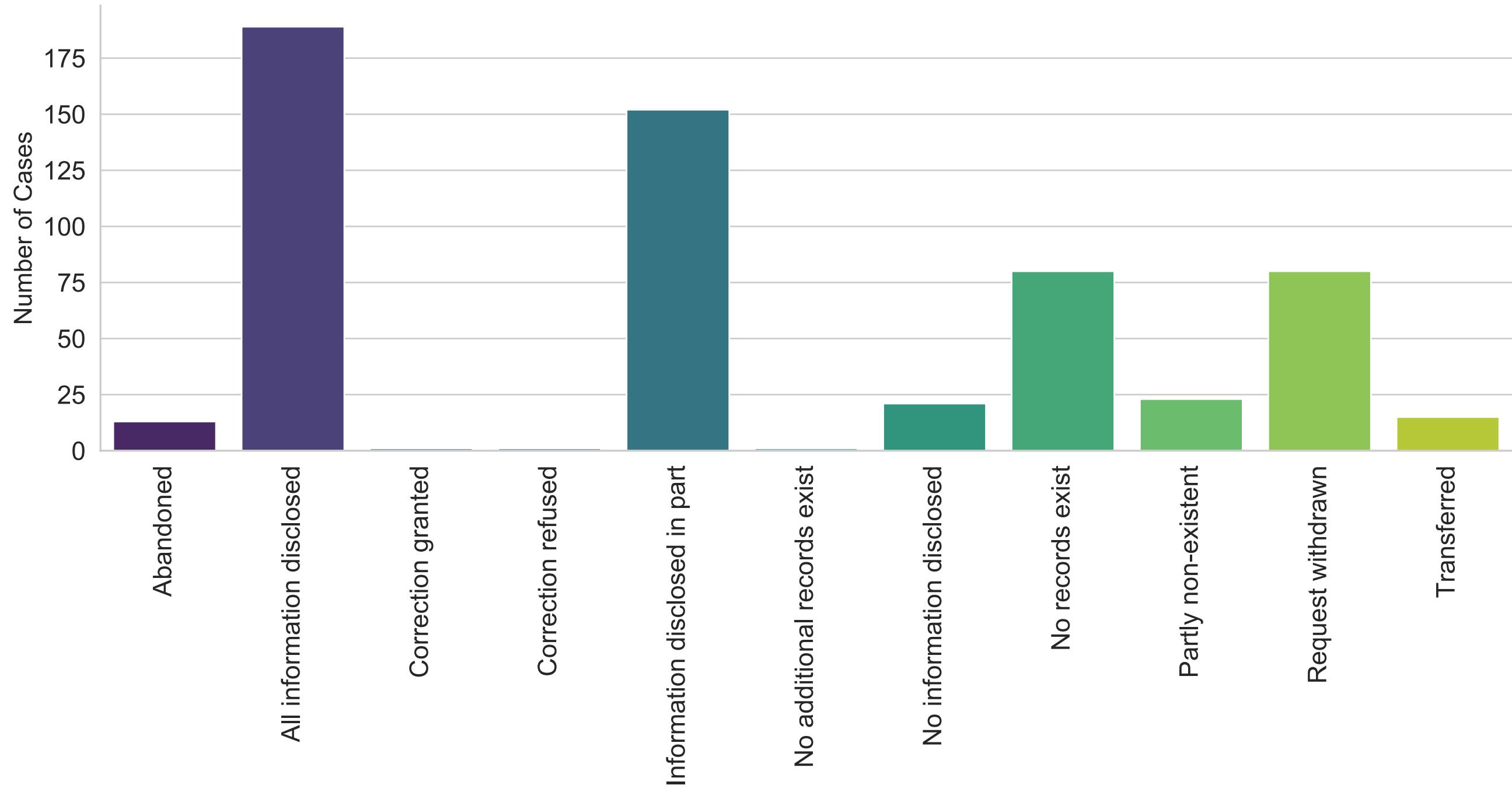
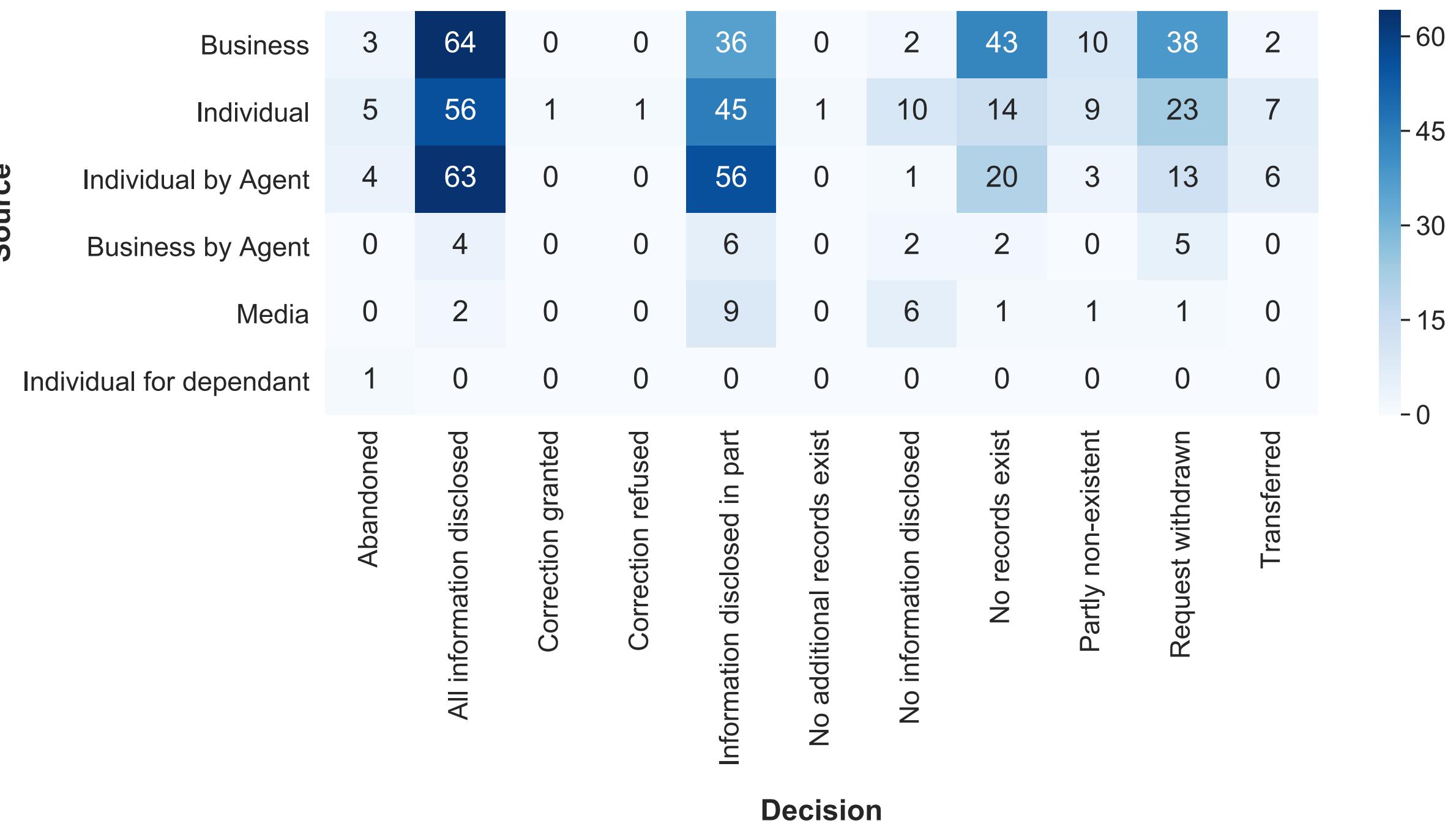


## Decisions Made for all Requests

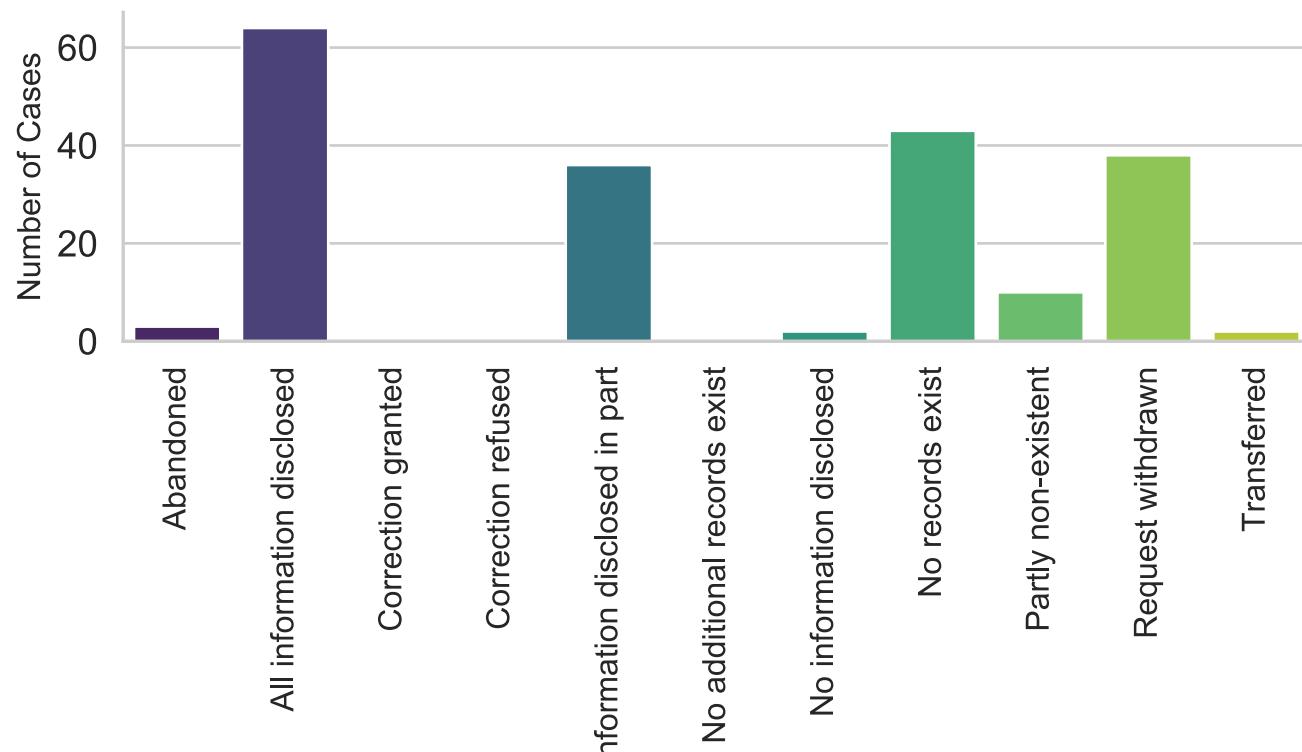


# Full Data

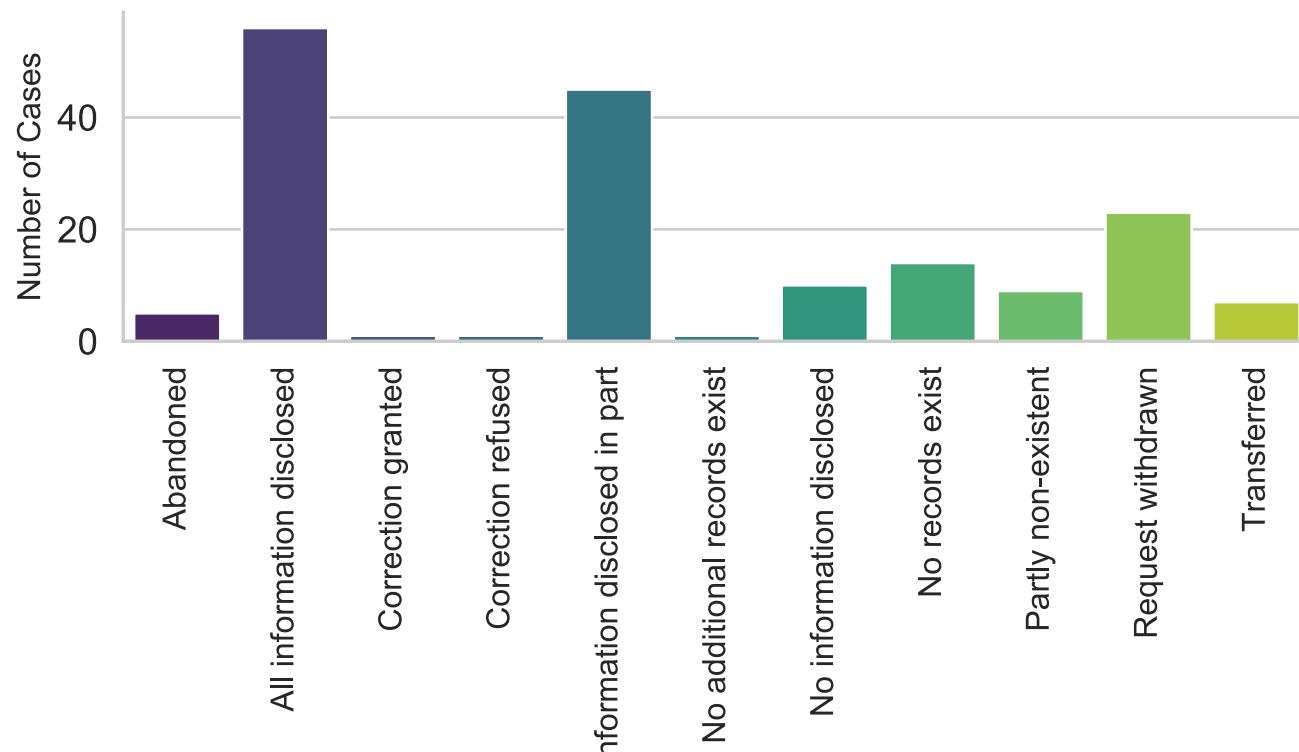


# Number of cases for all type of decisions made for each of the sources

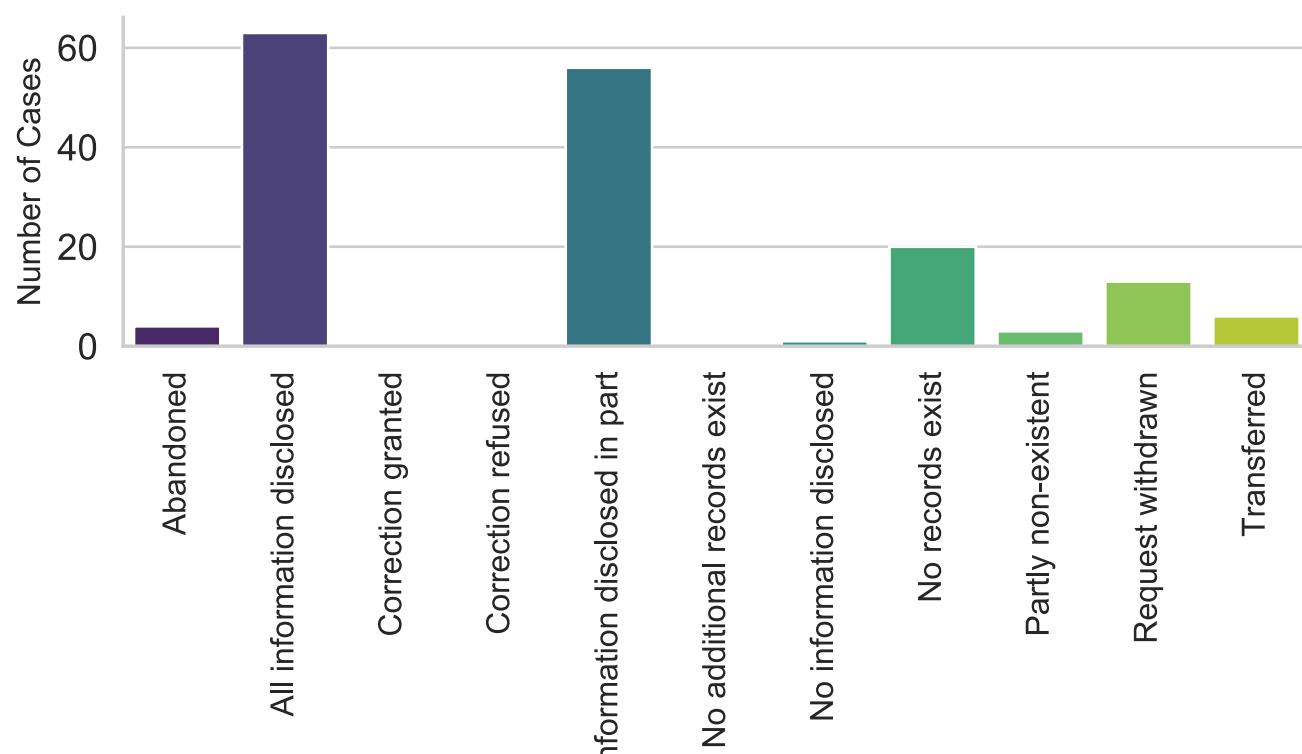
Requests made by 'Business'



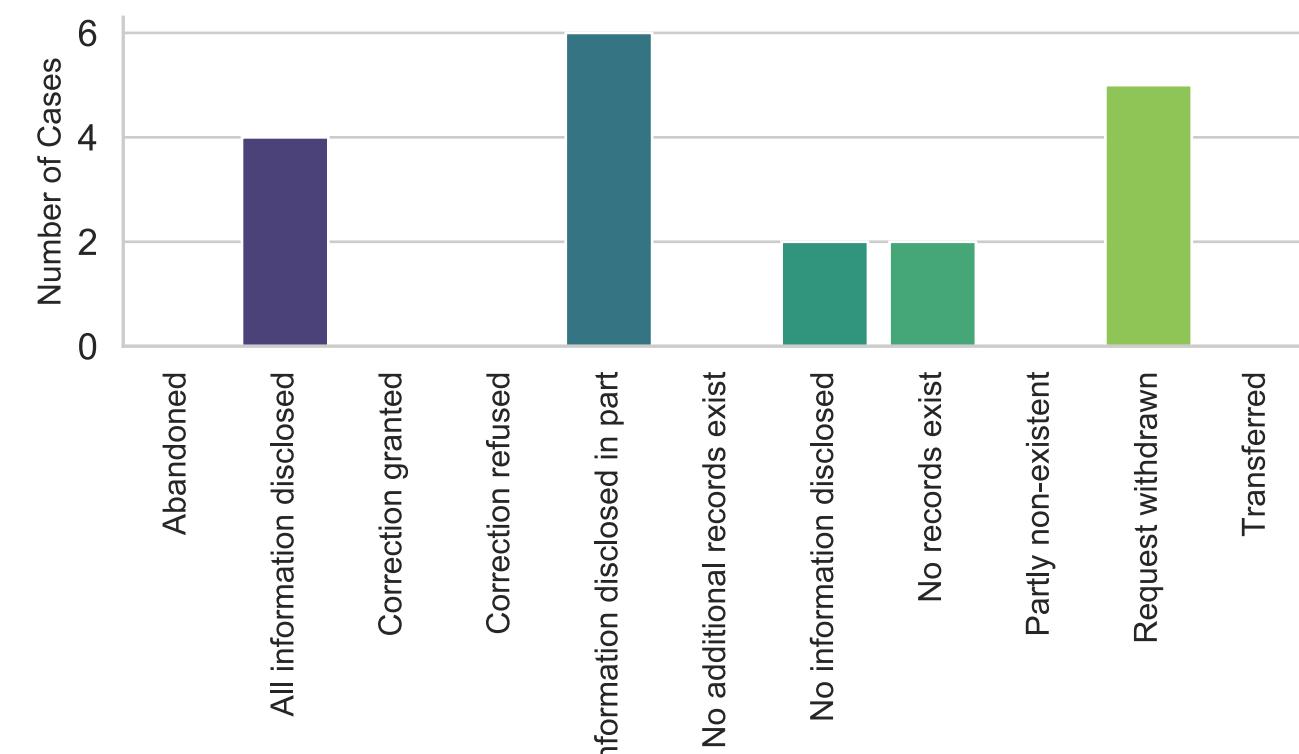
Requests made by 'Individual'



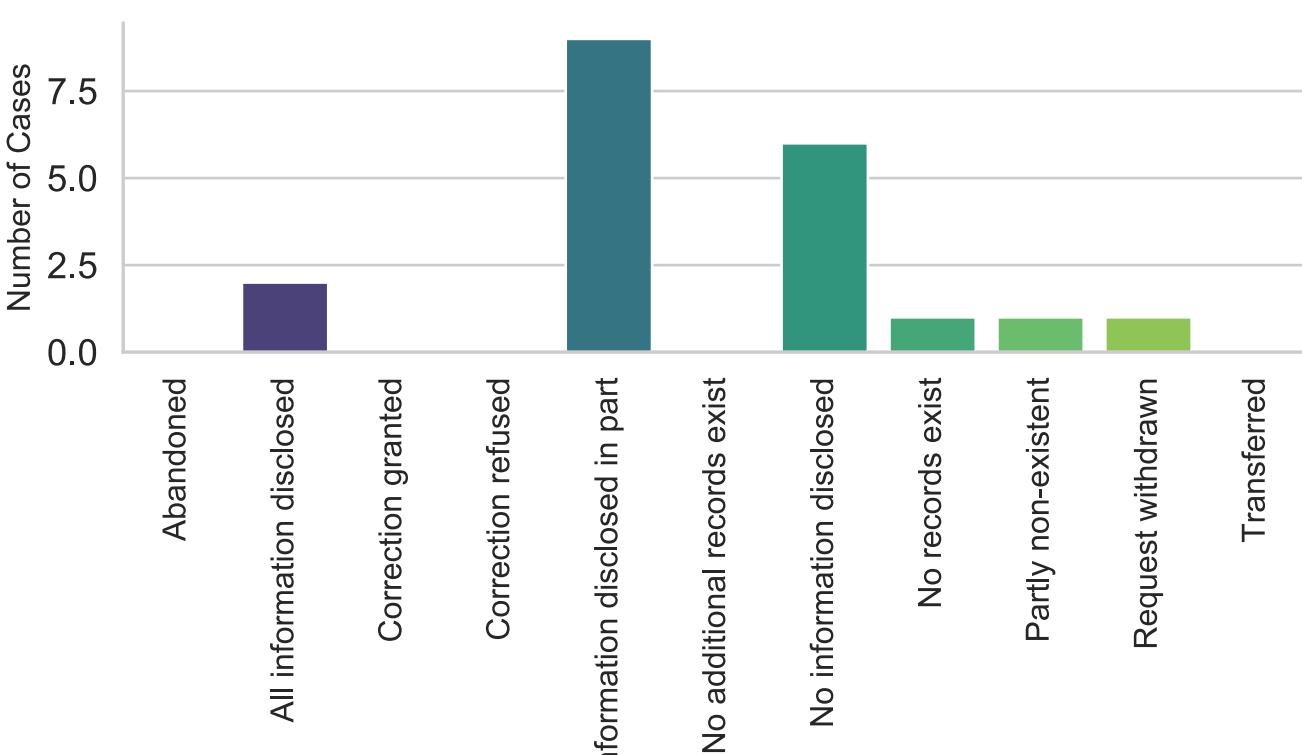
Requests made by 'Individual by Agent'



Requests made by 'Business by Agent'



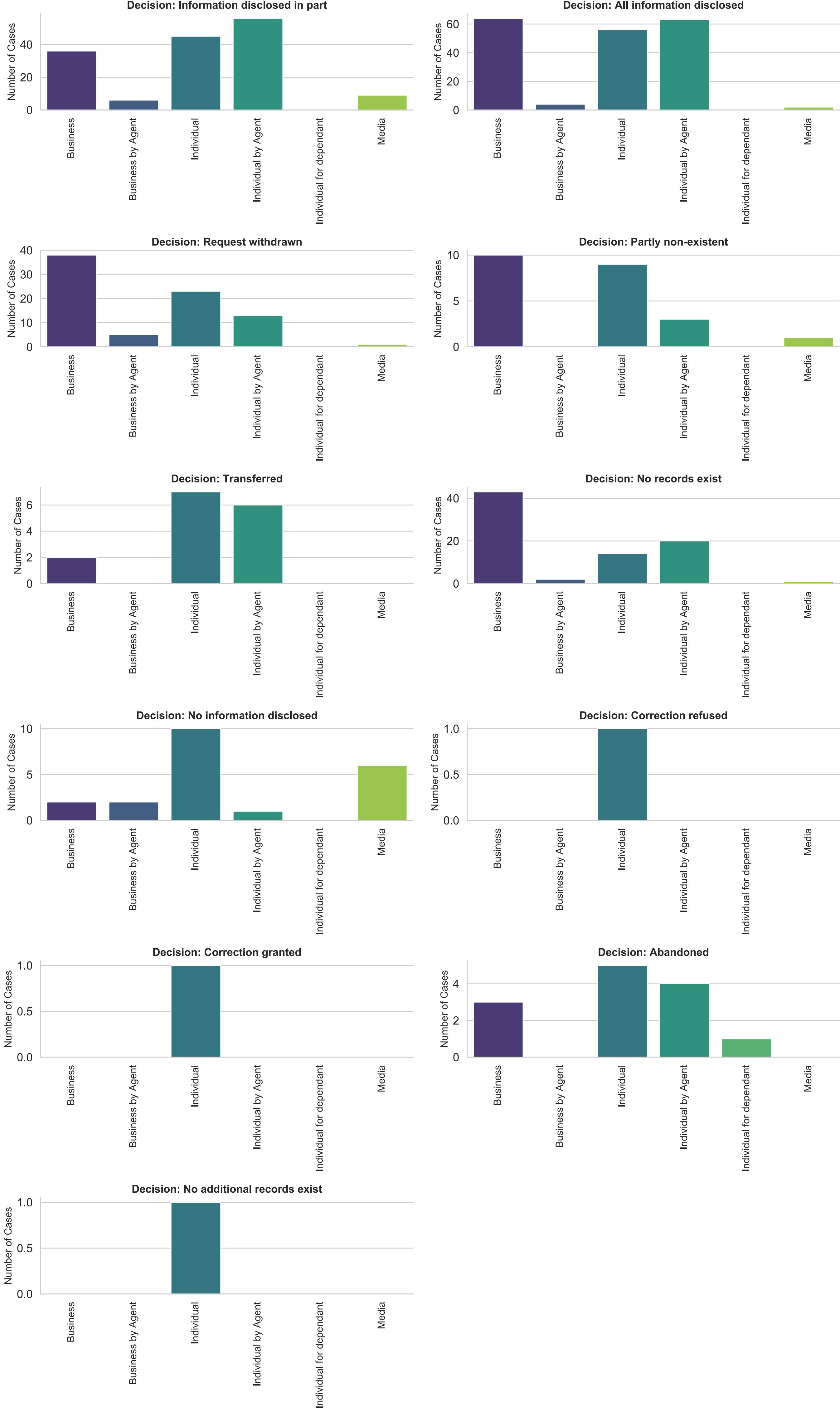
Requests made by 'Media'



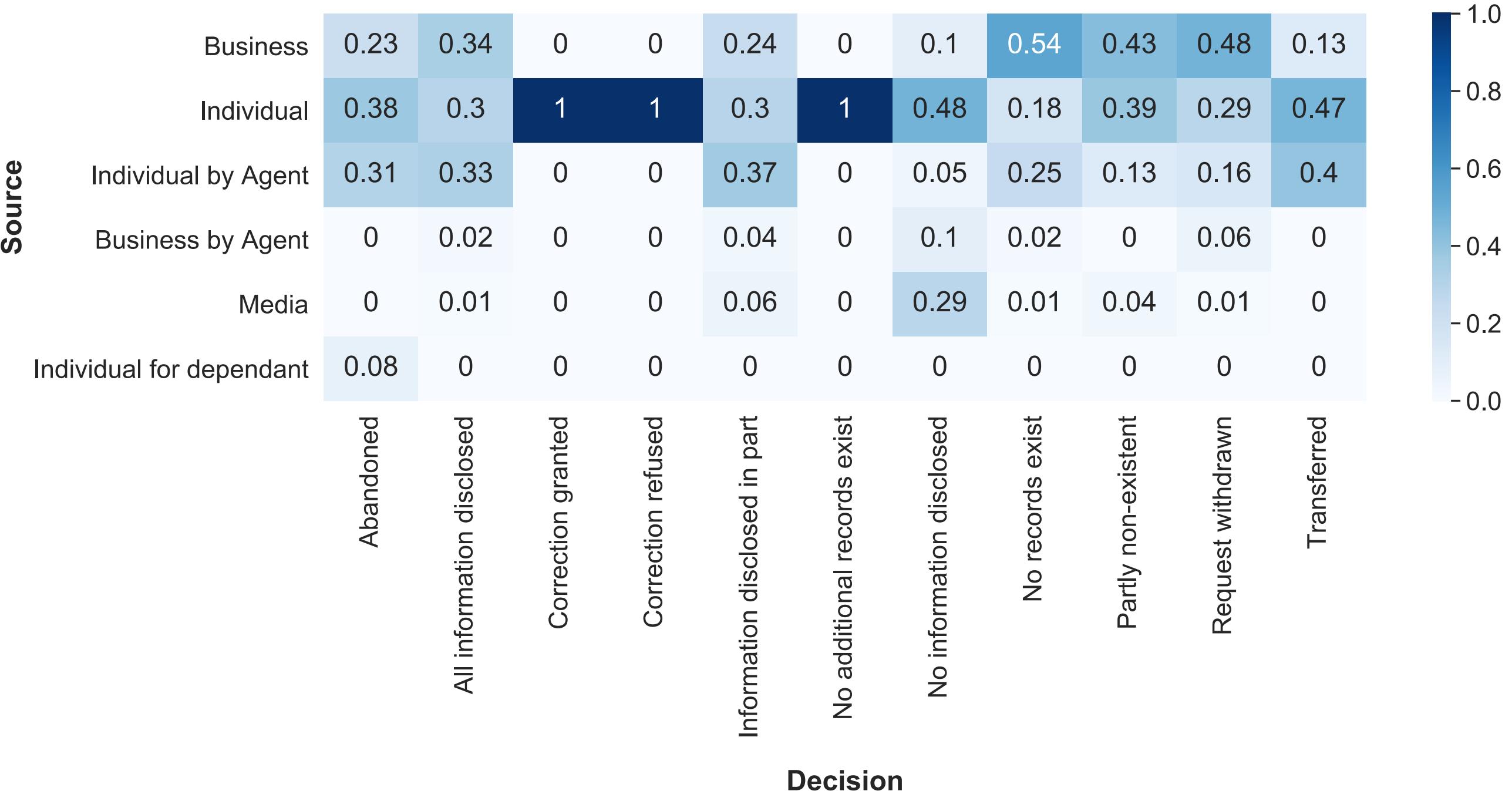
Requests made by 'Individual for dependant'



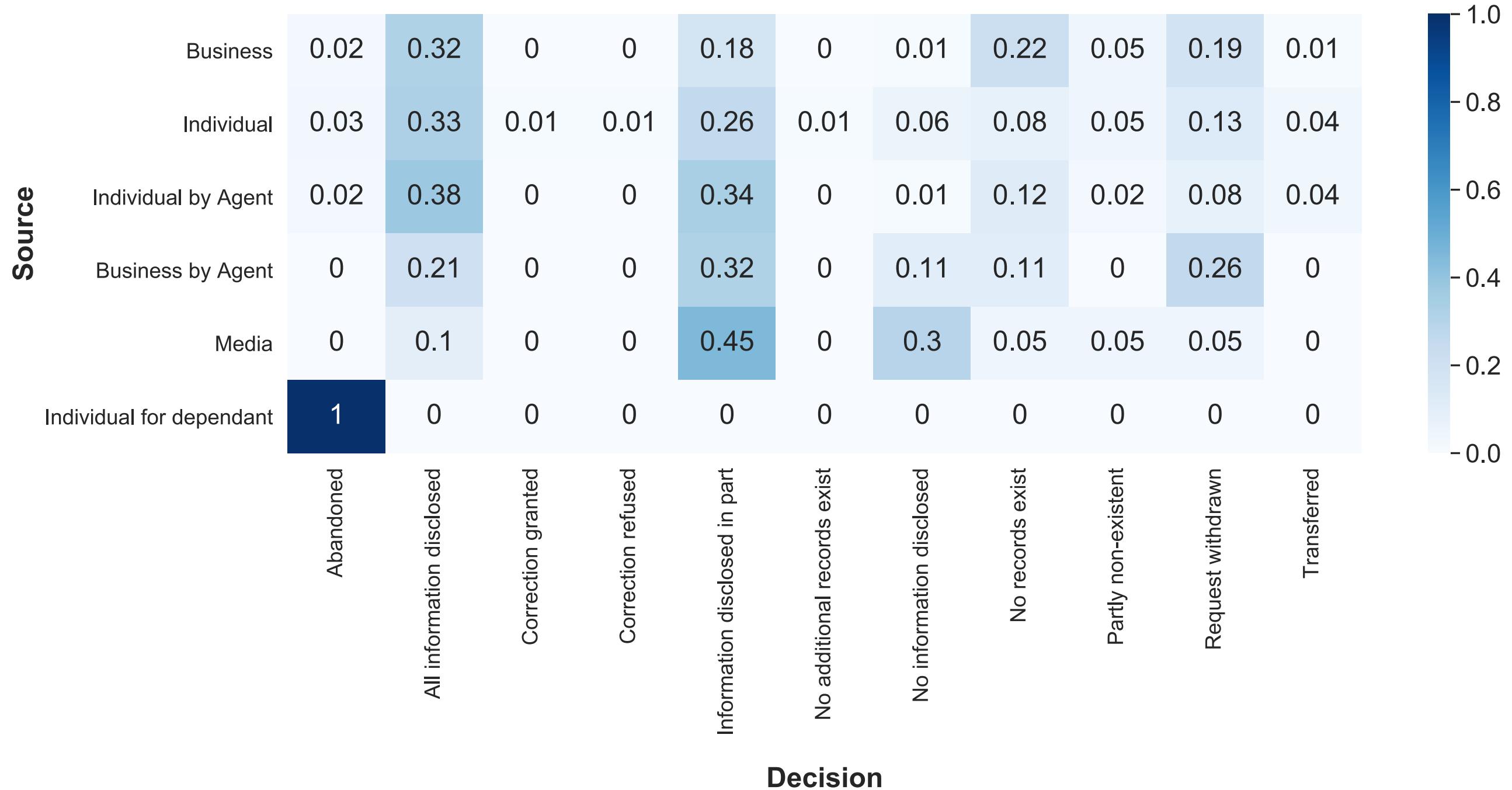
## Number of cases for each type of decision made by sources



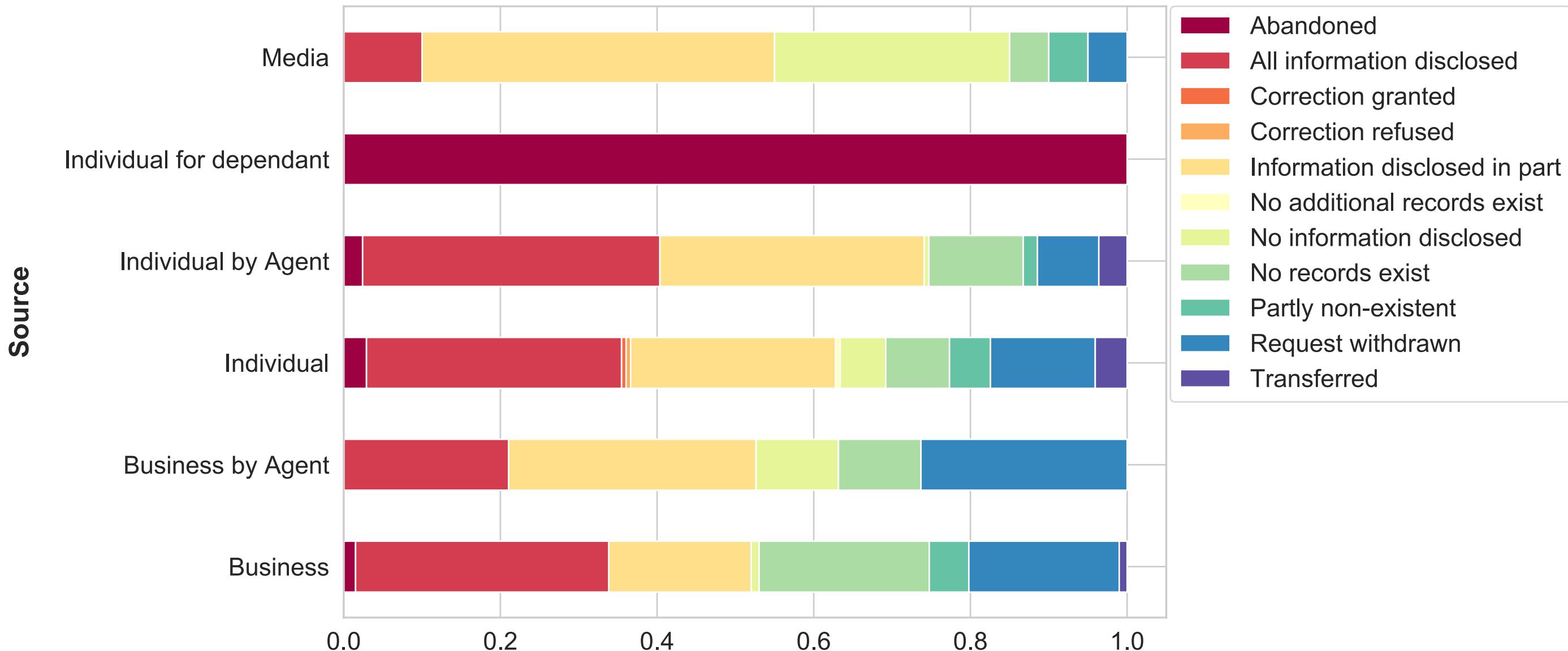
## How each decision is split among all sources (fraction)



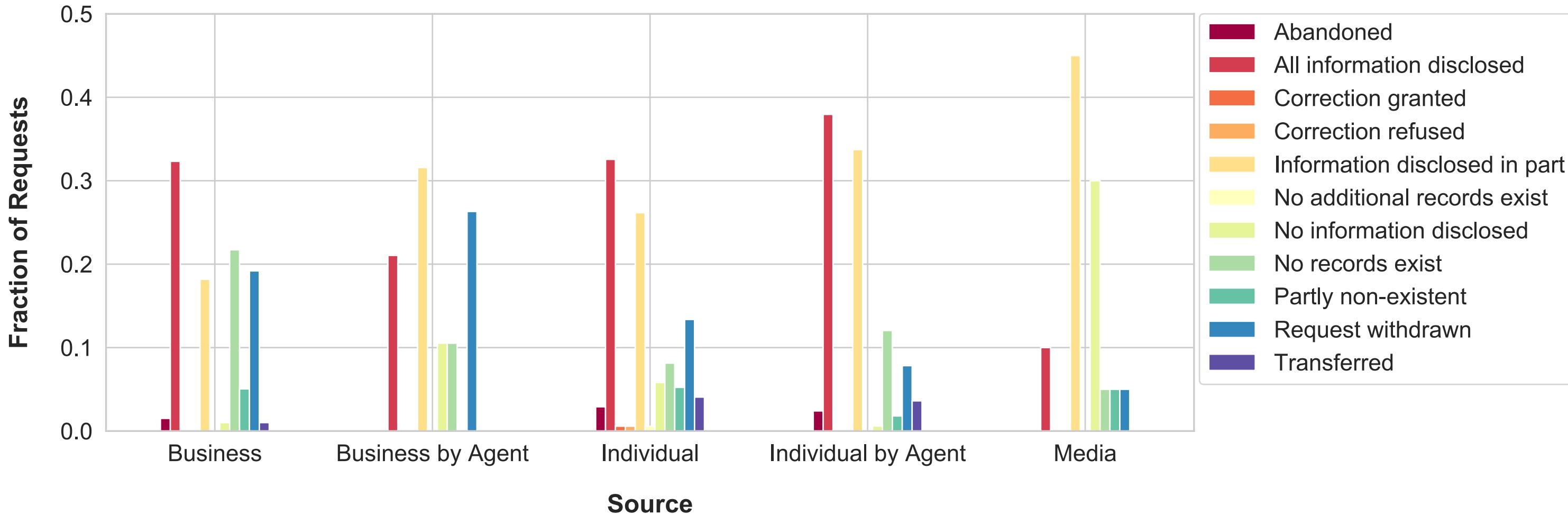
## How decisions are split per source (fraction)



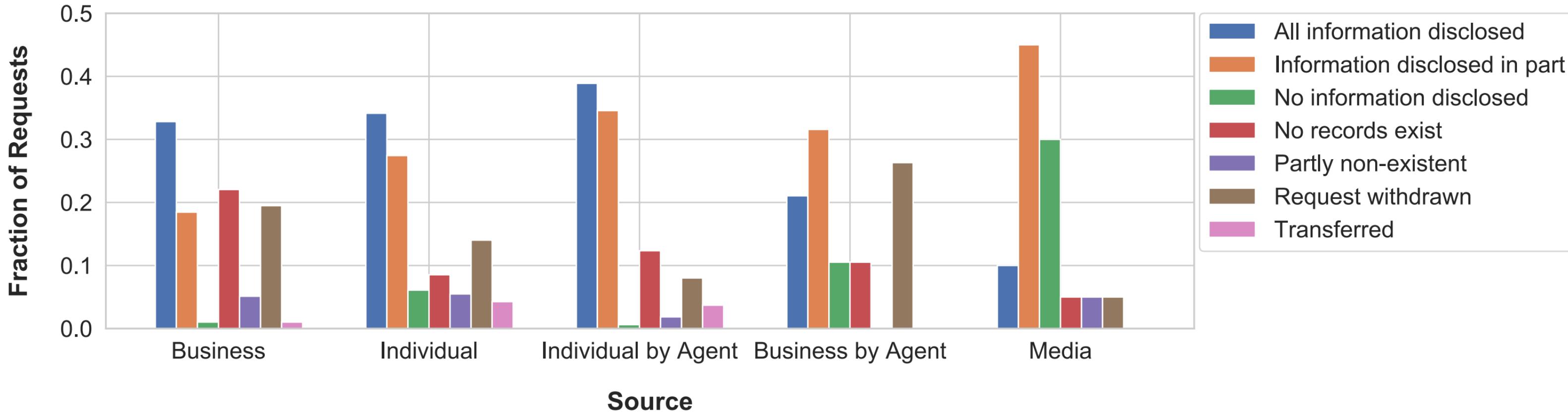
## Full data, fraction of decisions per source



## Full data, fraction of decisions per source



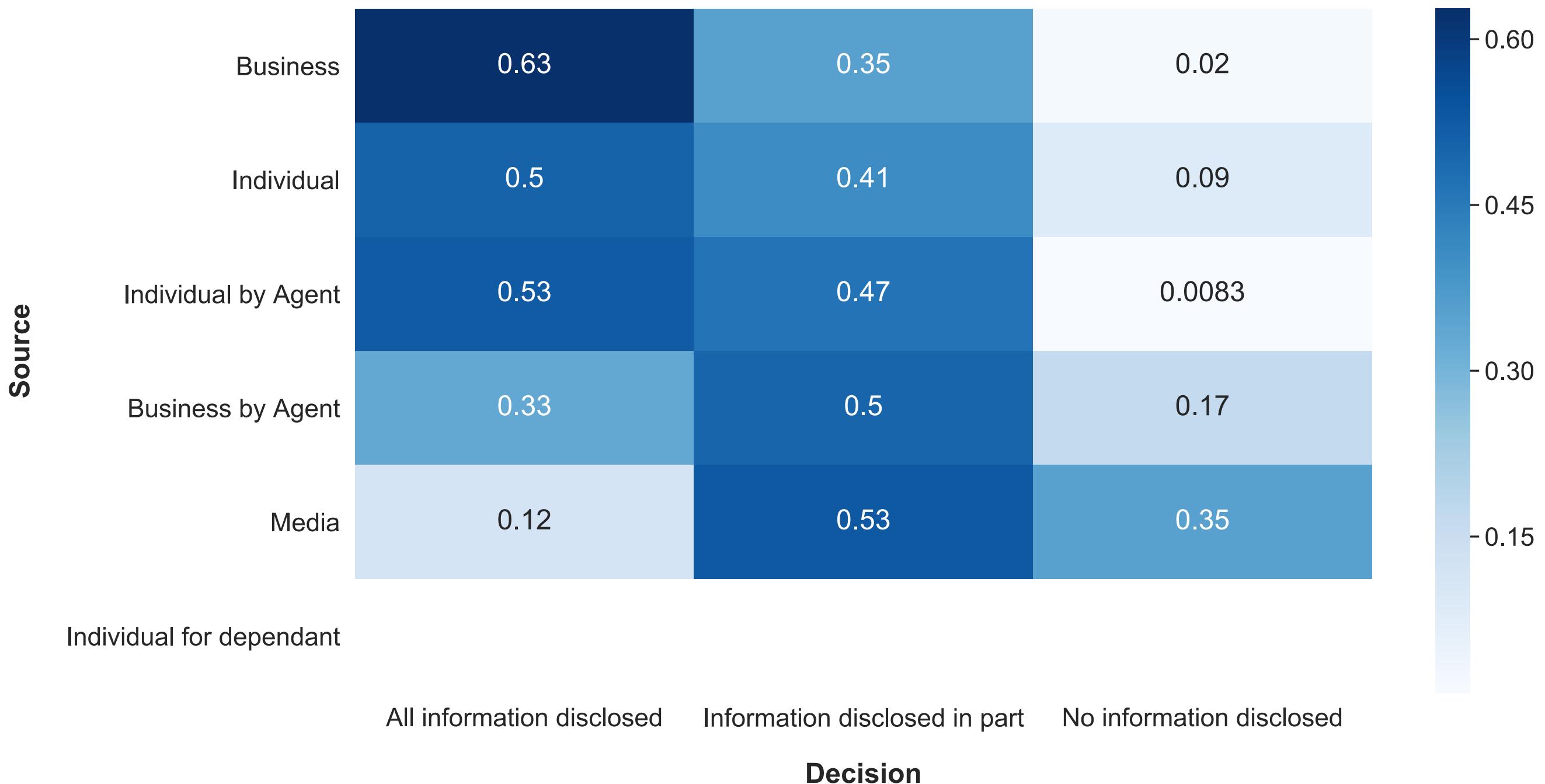
## Fraction of decisions per source, for decisions with more than 15 instances only



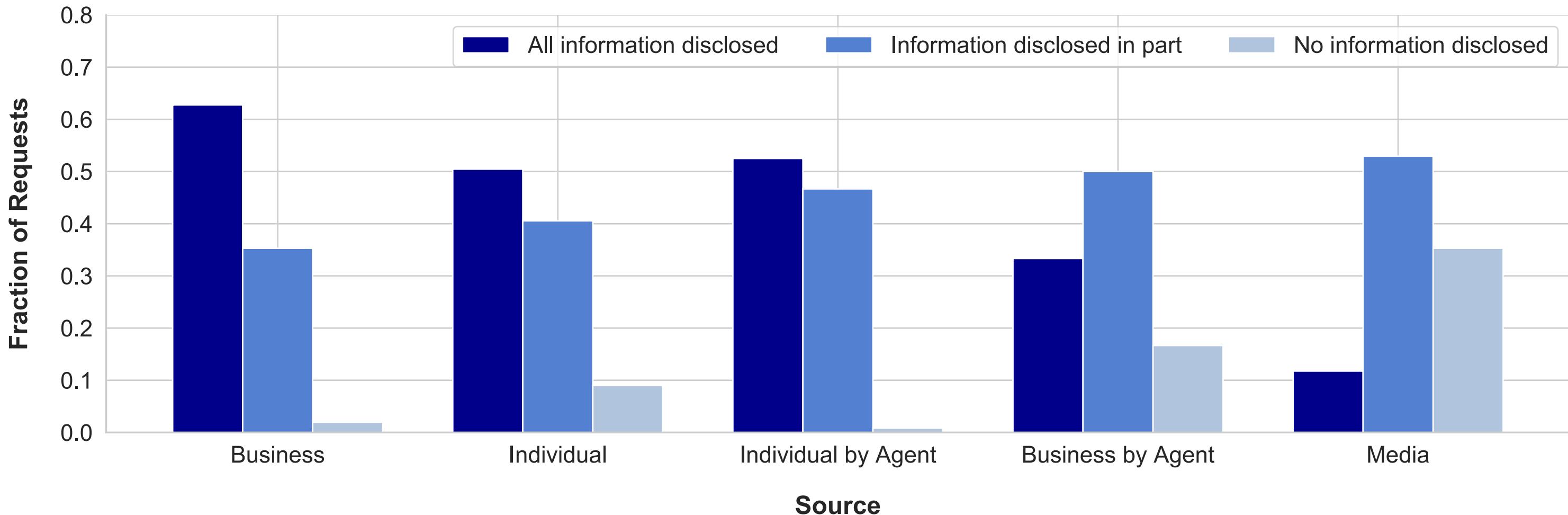
## Each of the three main decisions split among all the sources (fraction)



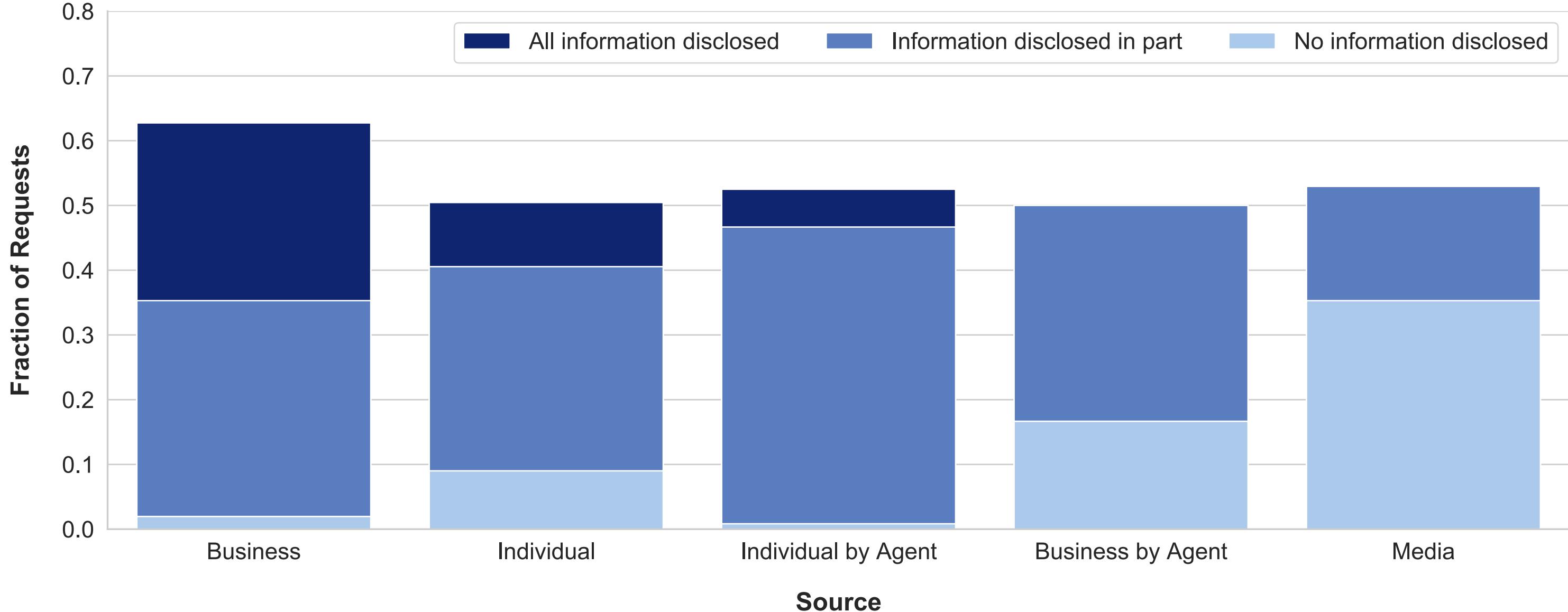
### Three main decisions split for each source (fraction)



## Three main decisions only, fractions for each source add to 1



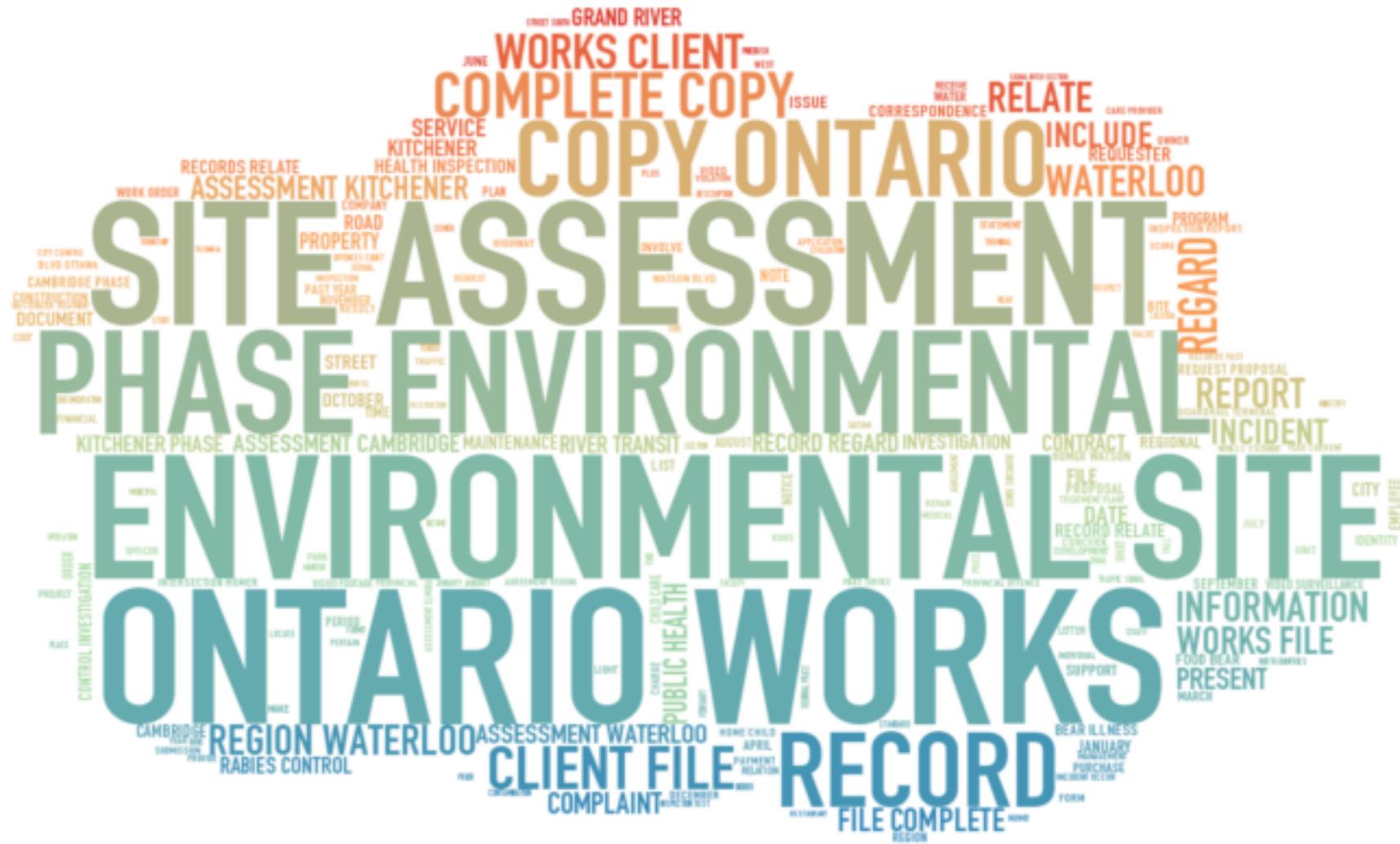
## Three main decisions only, fractions for each source add to 1



# Top 200 unigrams/bigrams, full text



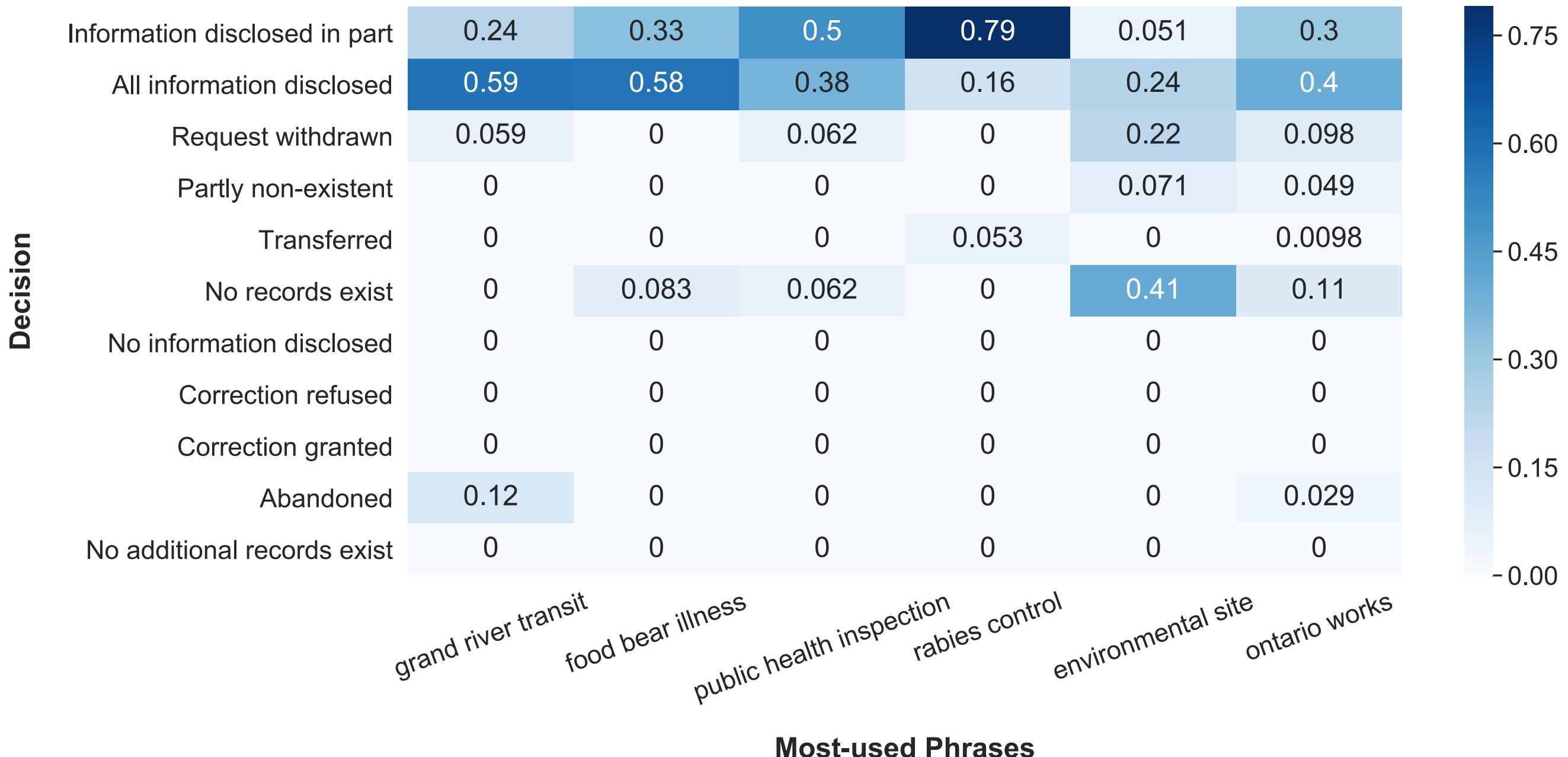
## Top 200 unigrams, full text



## Top 200 unigrams/bigrams, full text without '{\* remove}'

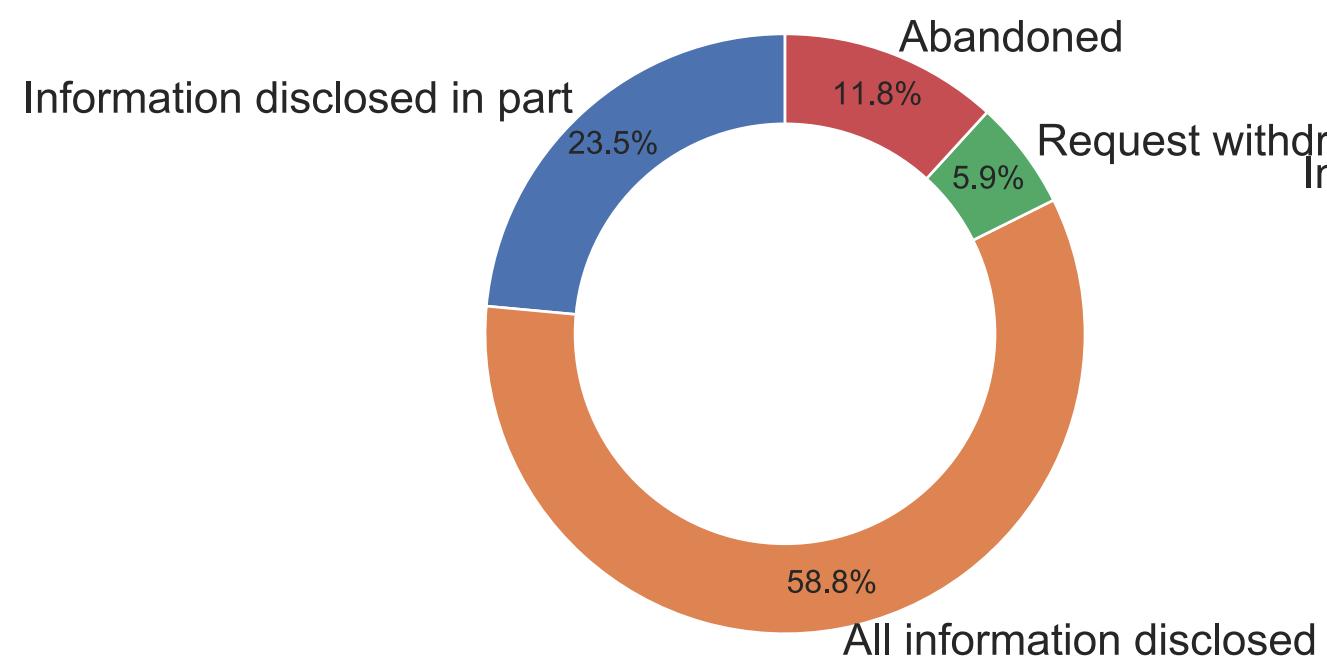
## Top 200 unigrams, full text without '{\* remove}'

**46% of the full data uses the following phrases.**  
**For each phrase, here is how decisions are split (fraction).**

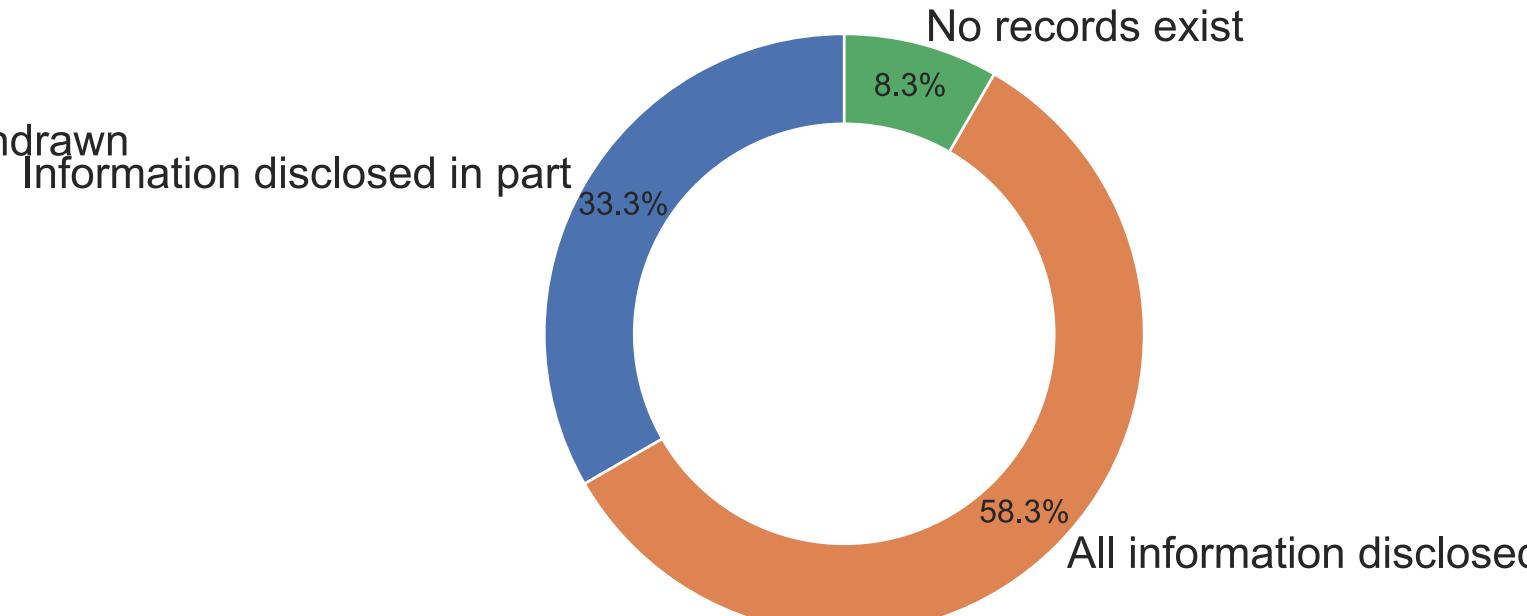


## Decision percentage for each n-gram

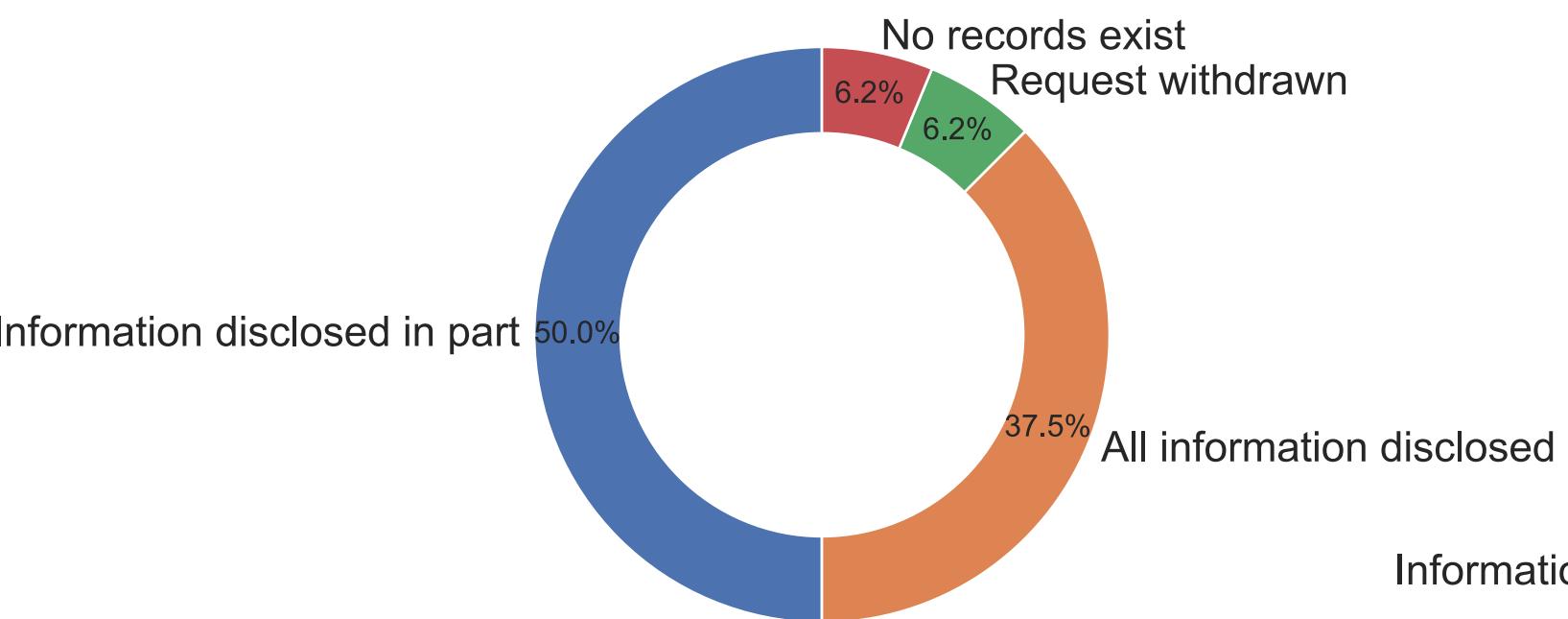
**grand river transit**



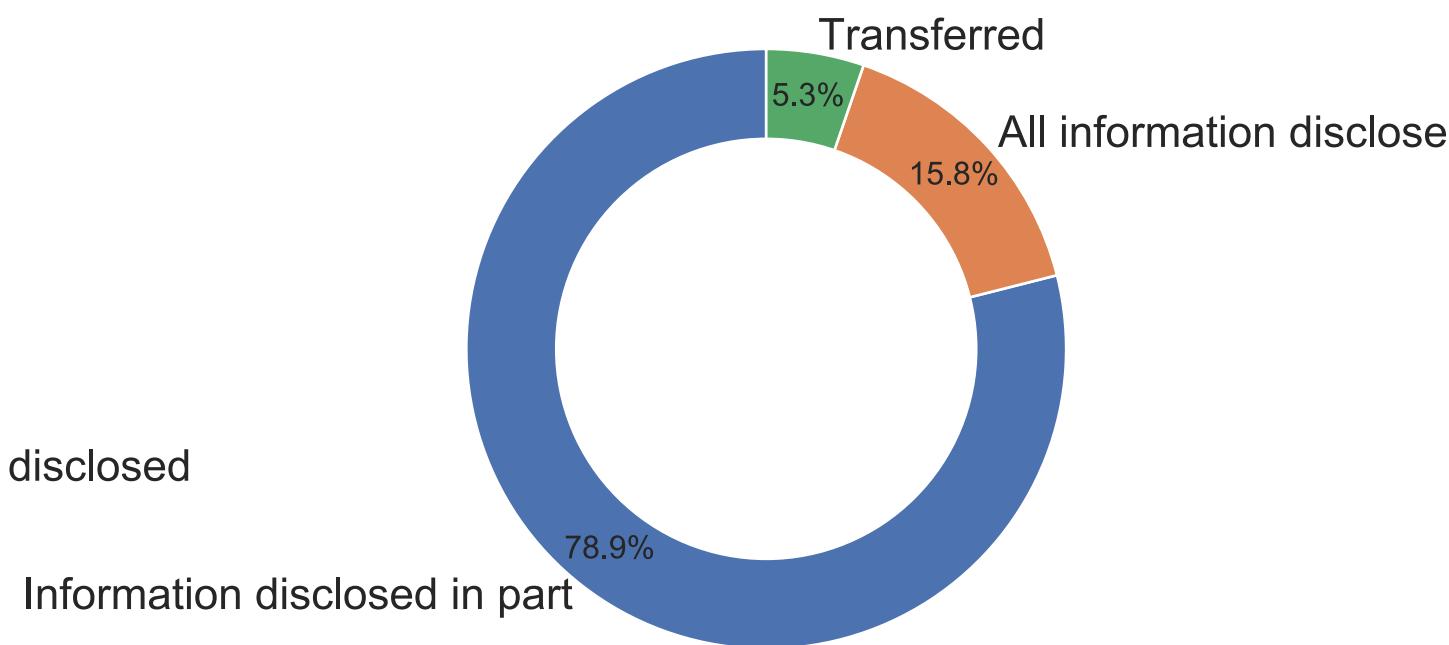
**food bear illness**



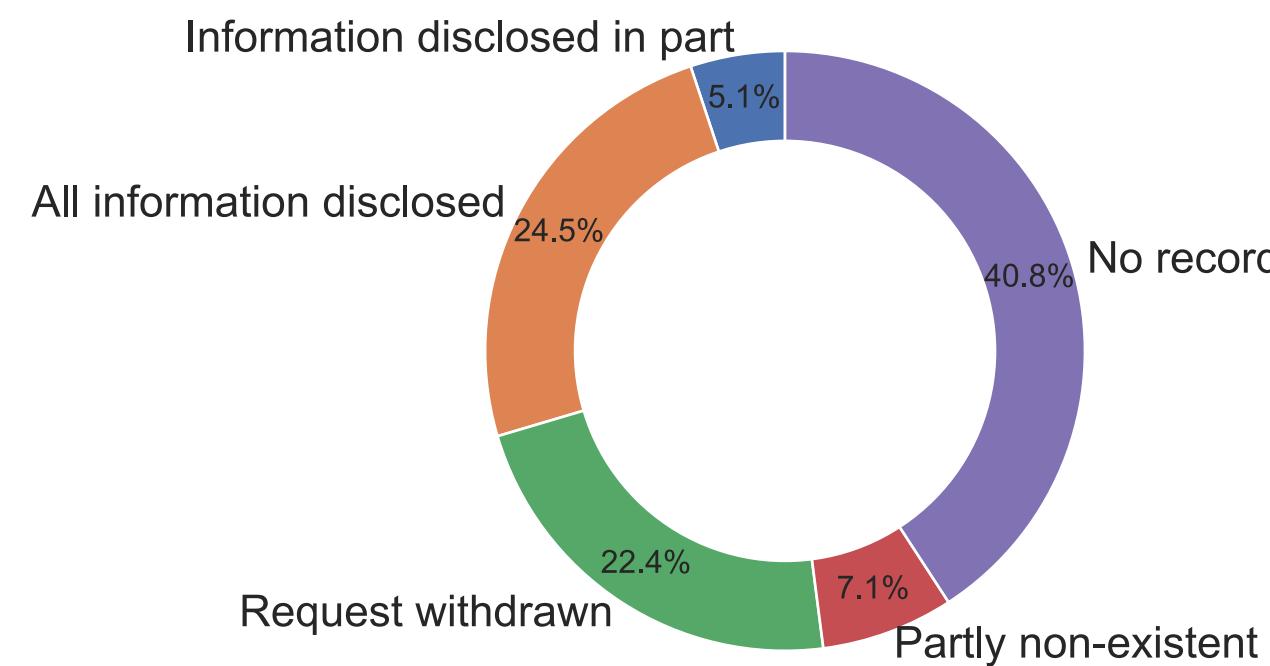
**public health inspection**



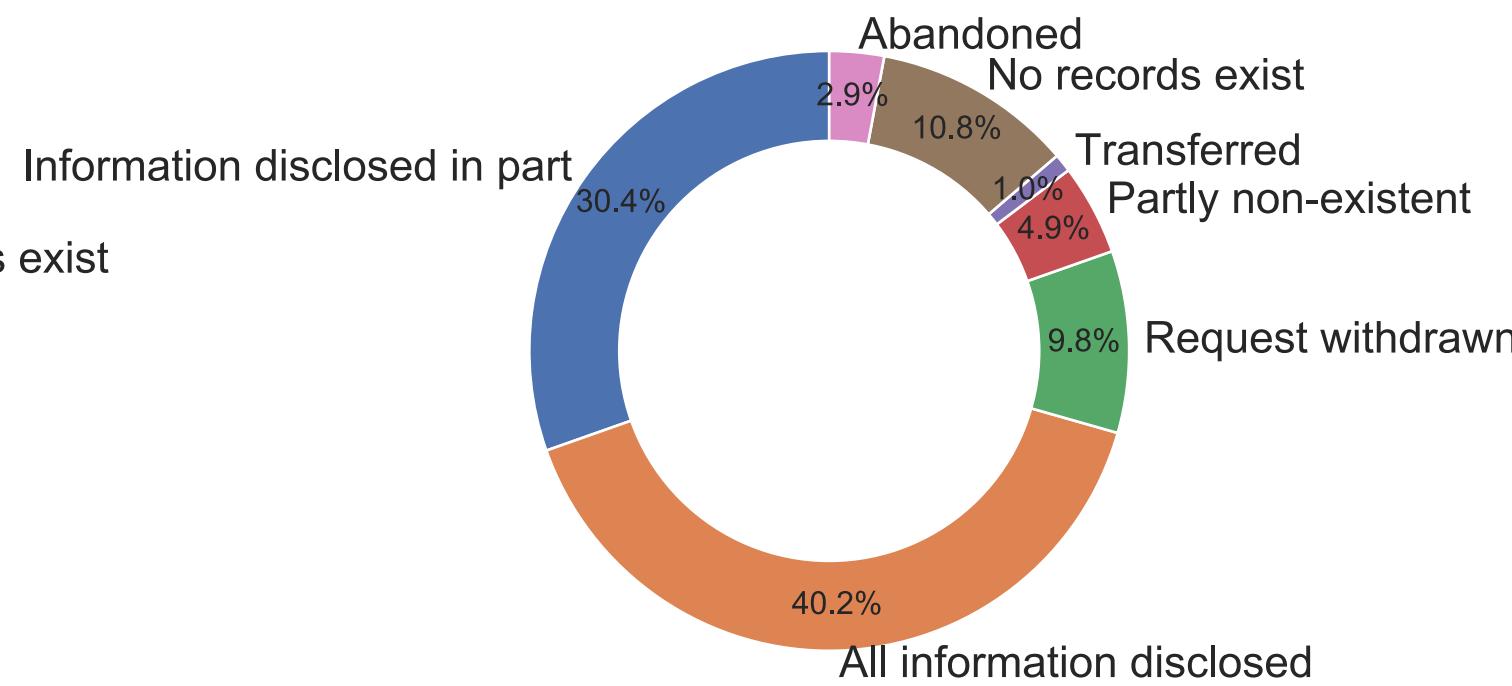
**rabies control**



**environmental site**

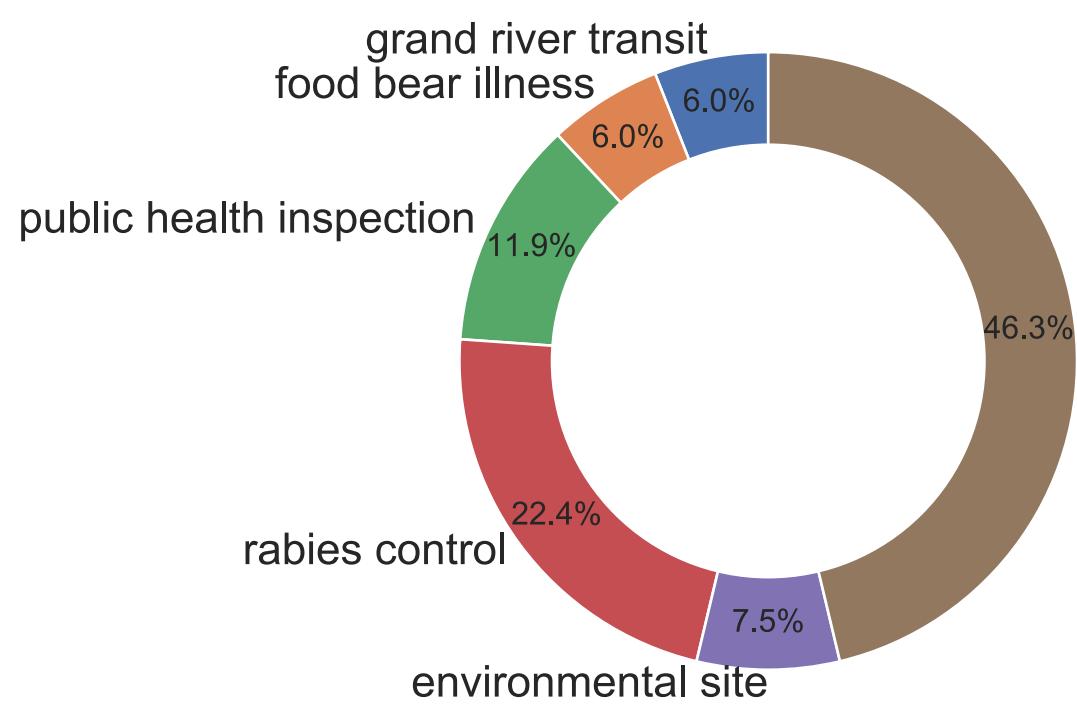


**ontario works**

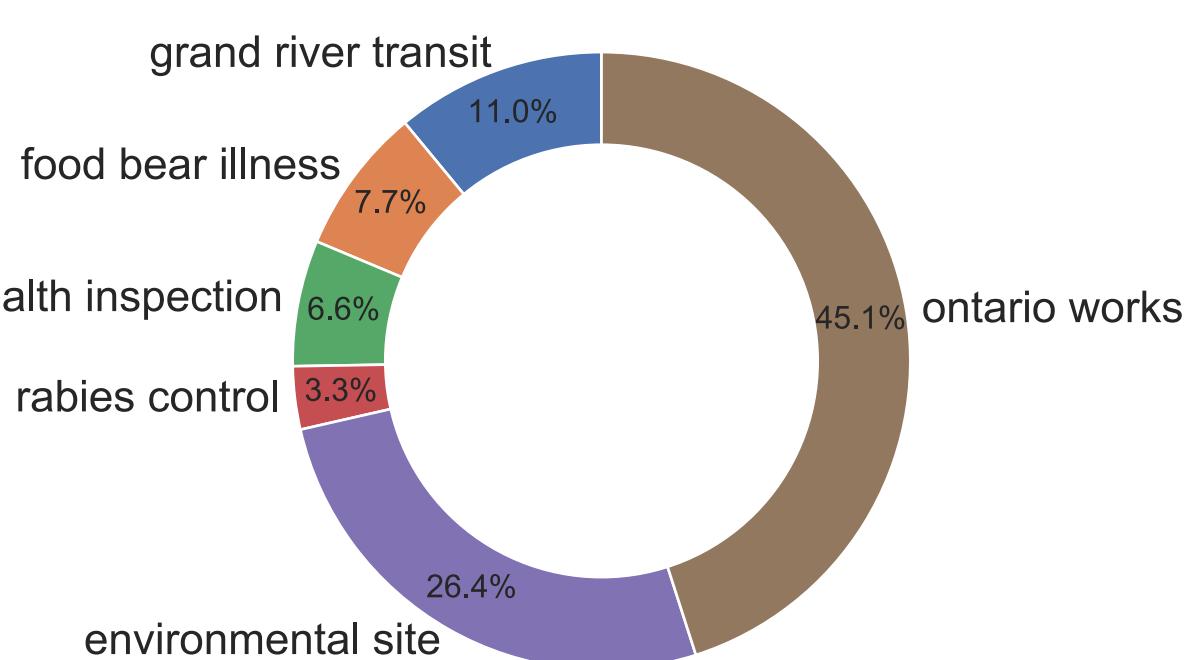


## For requests with the n-grams, n-gram percentage based on decision

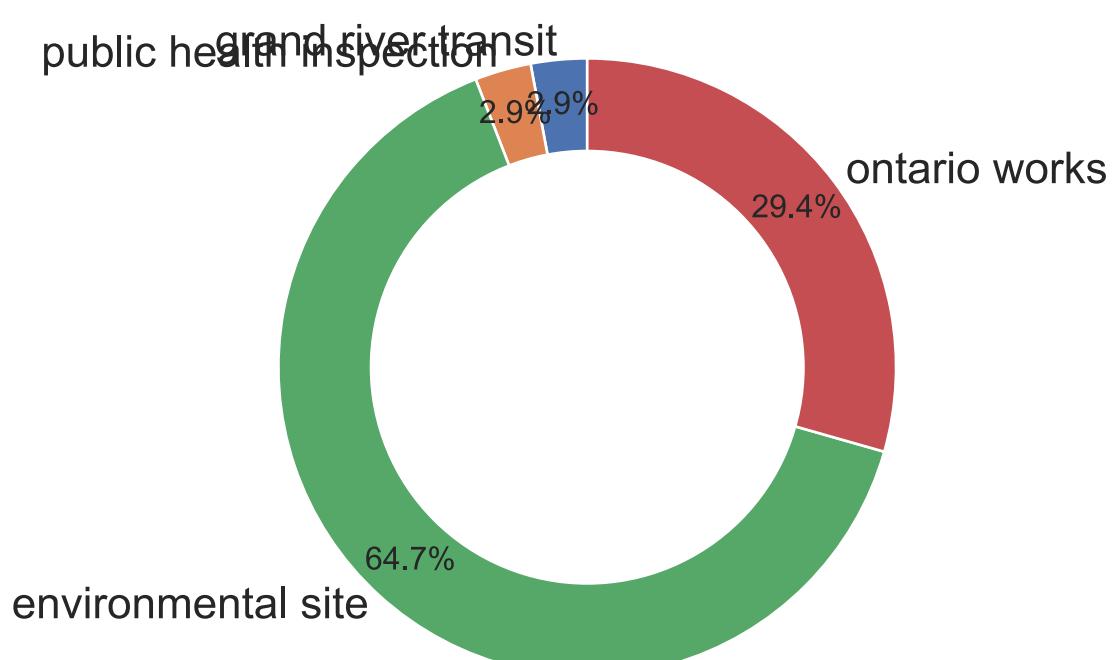
### Information disclosed in part



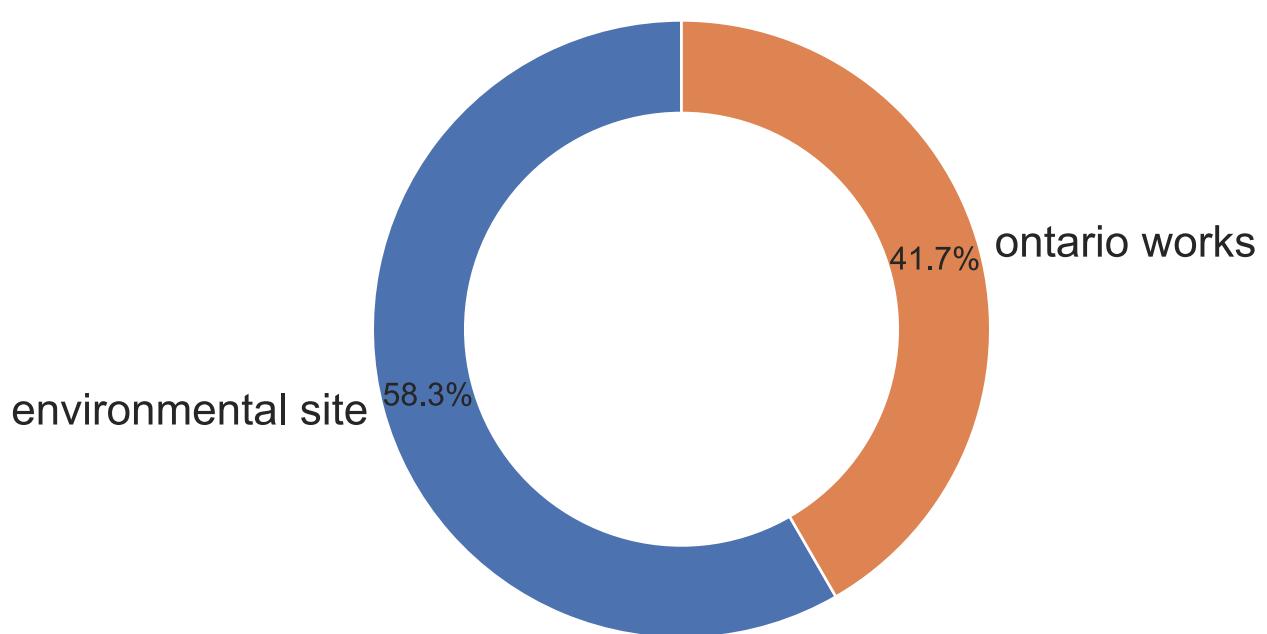
### All information disclosed



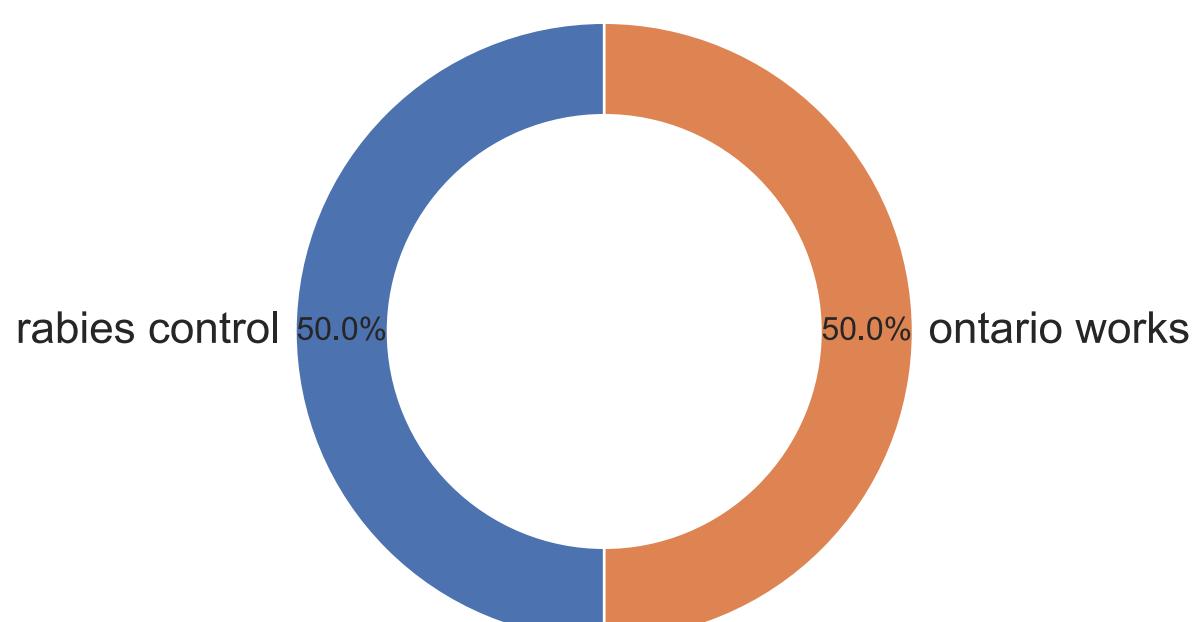
### Request withdrawn



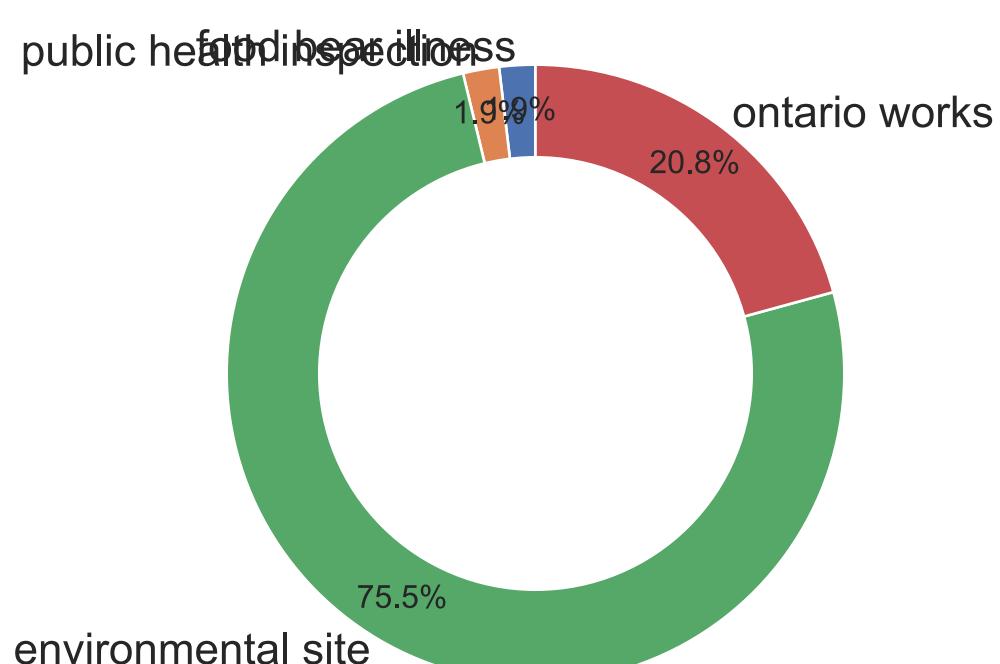
### Partly non-existent



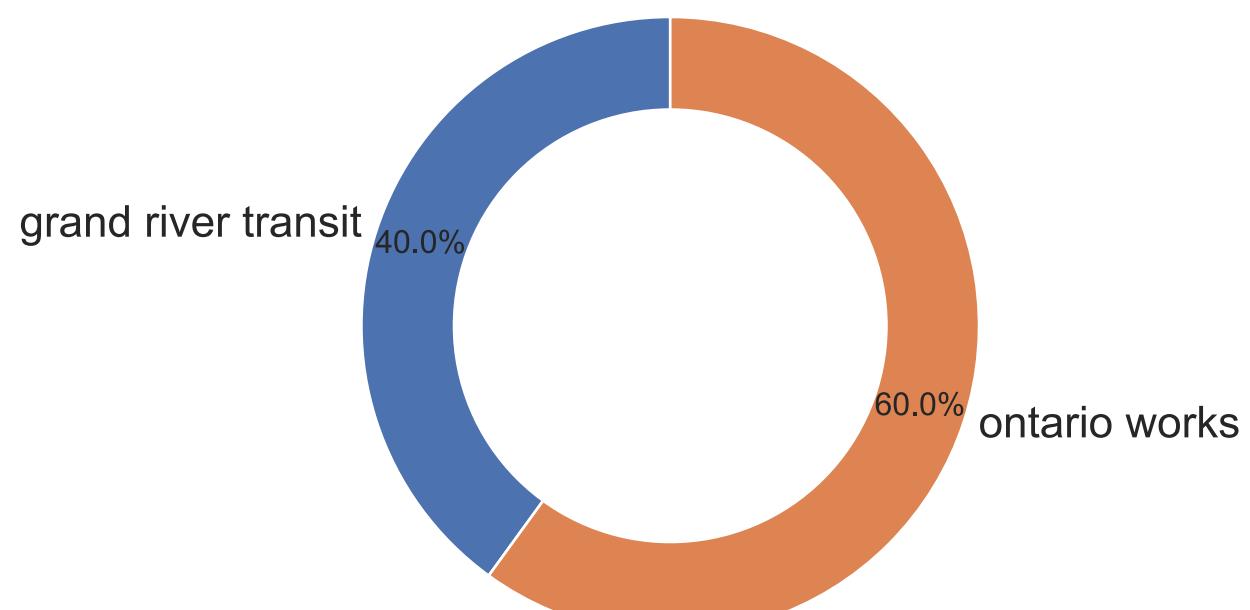
### Transferred



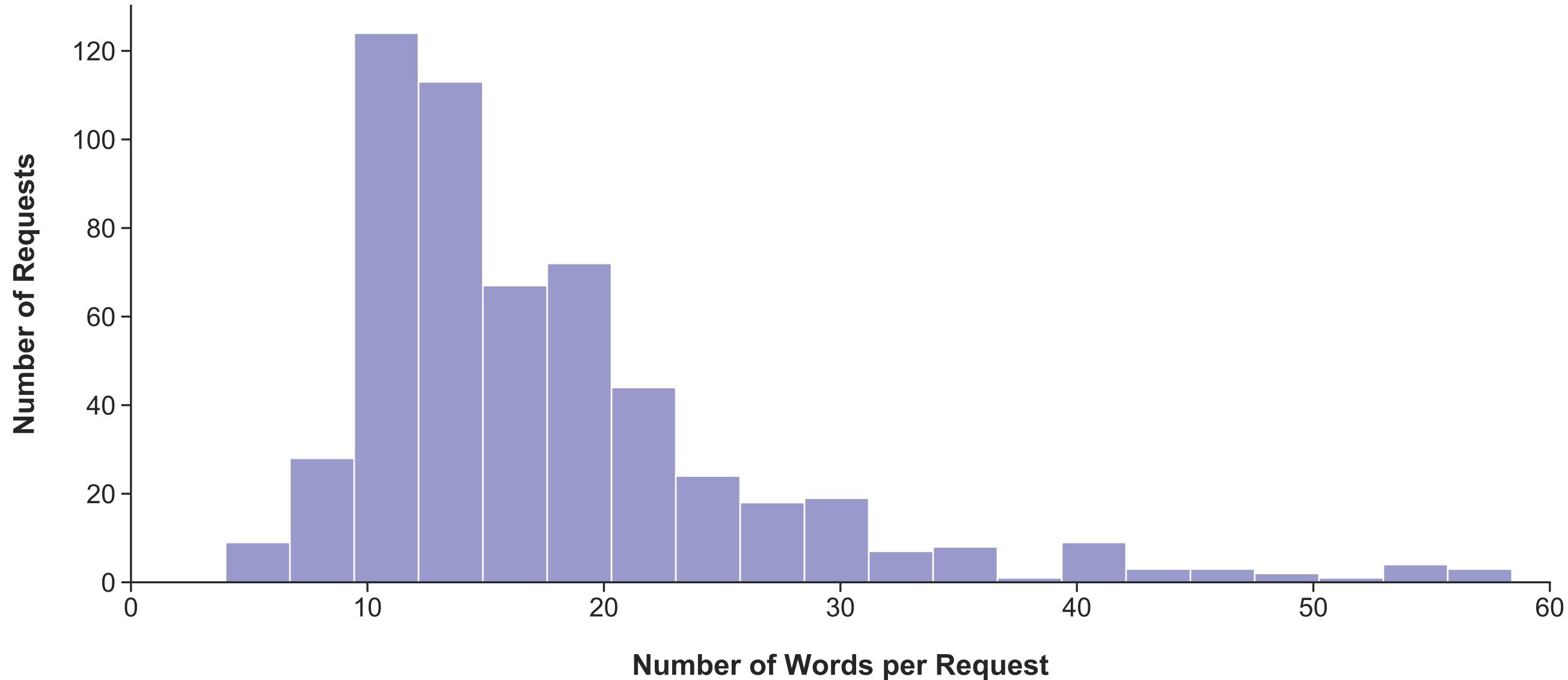
### No records exist



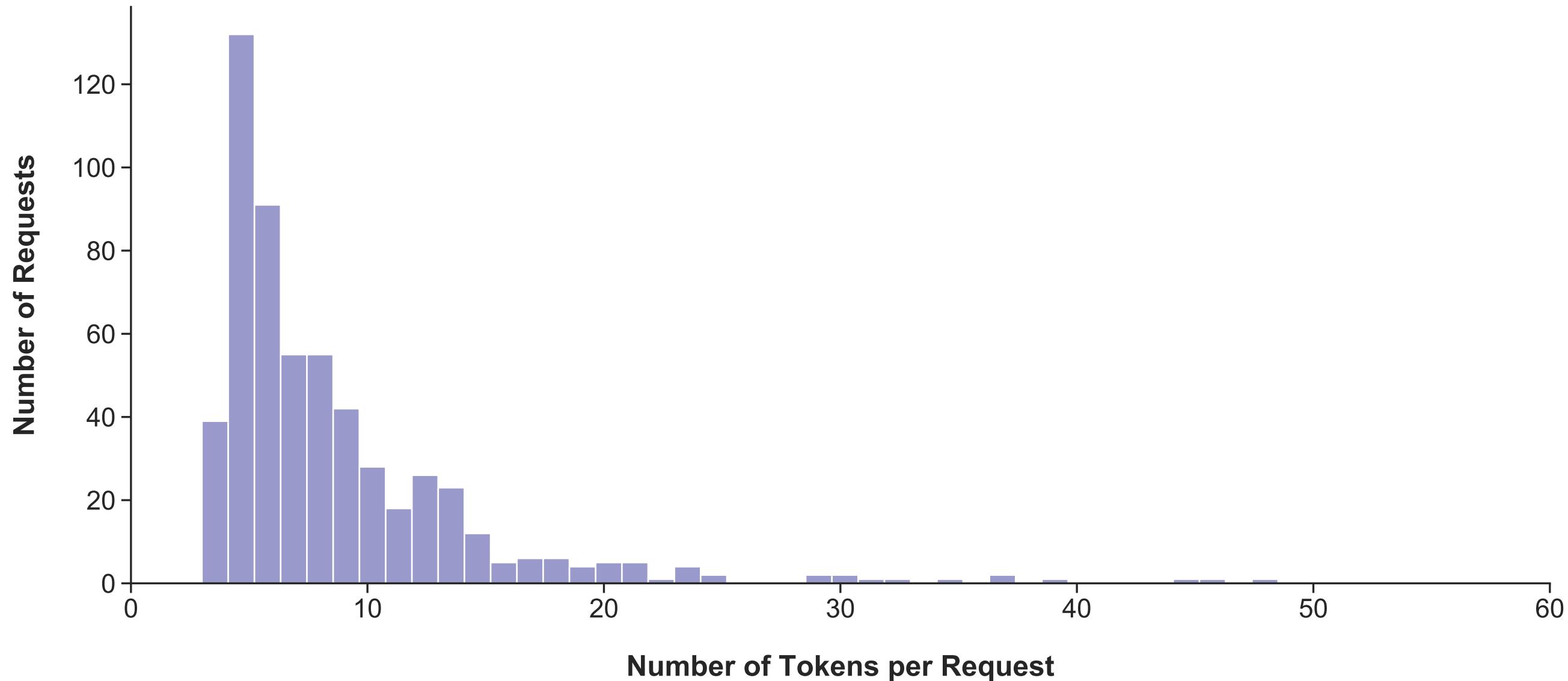
### Abandoned



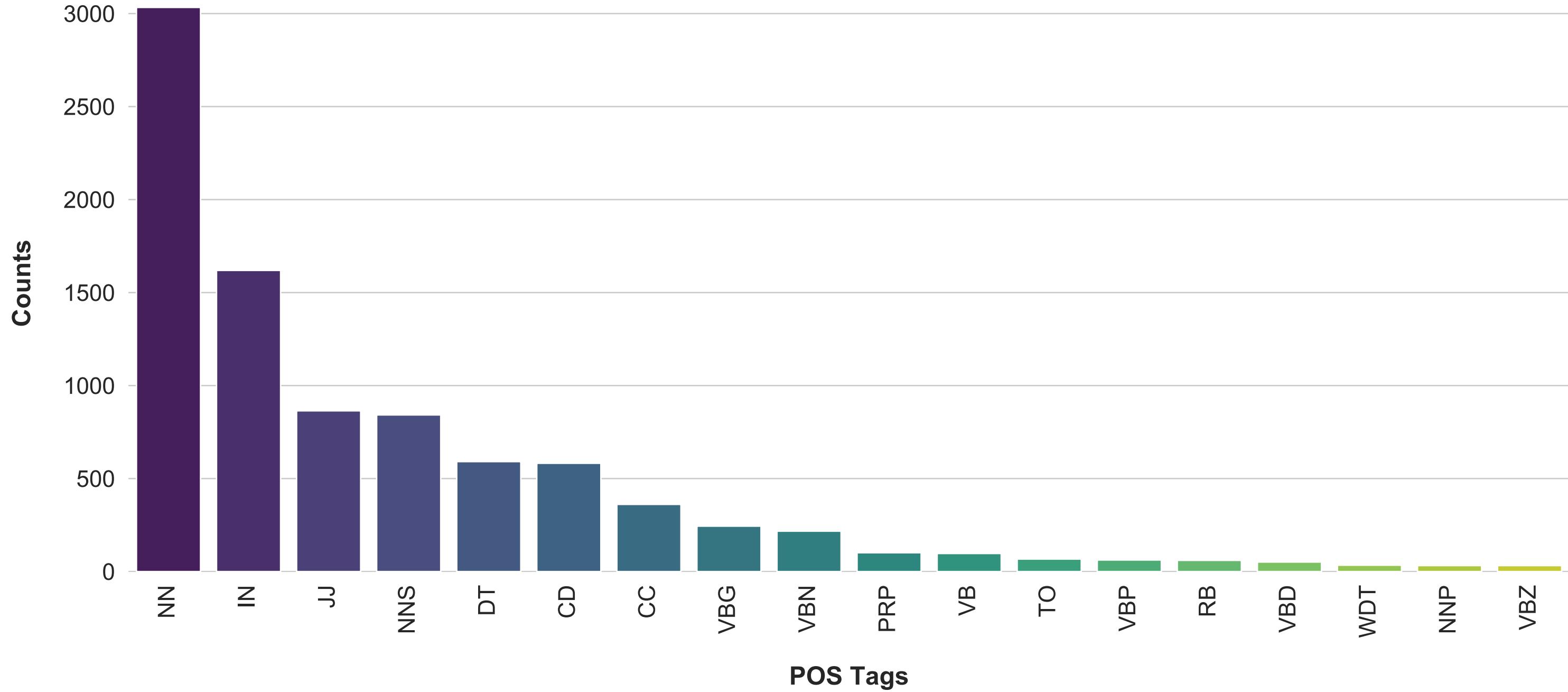
**Average number of words per request is 20.5, while the median is 15.0**



**Average number of tokens per request is 9.3, while the median is 7.0**

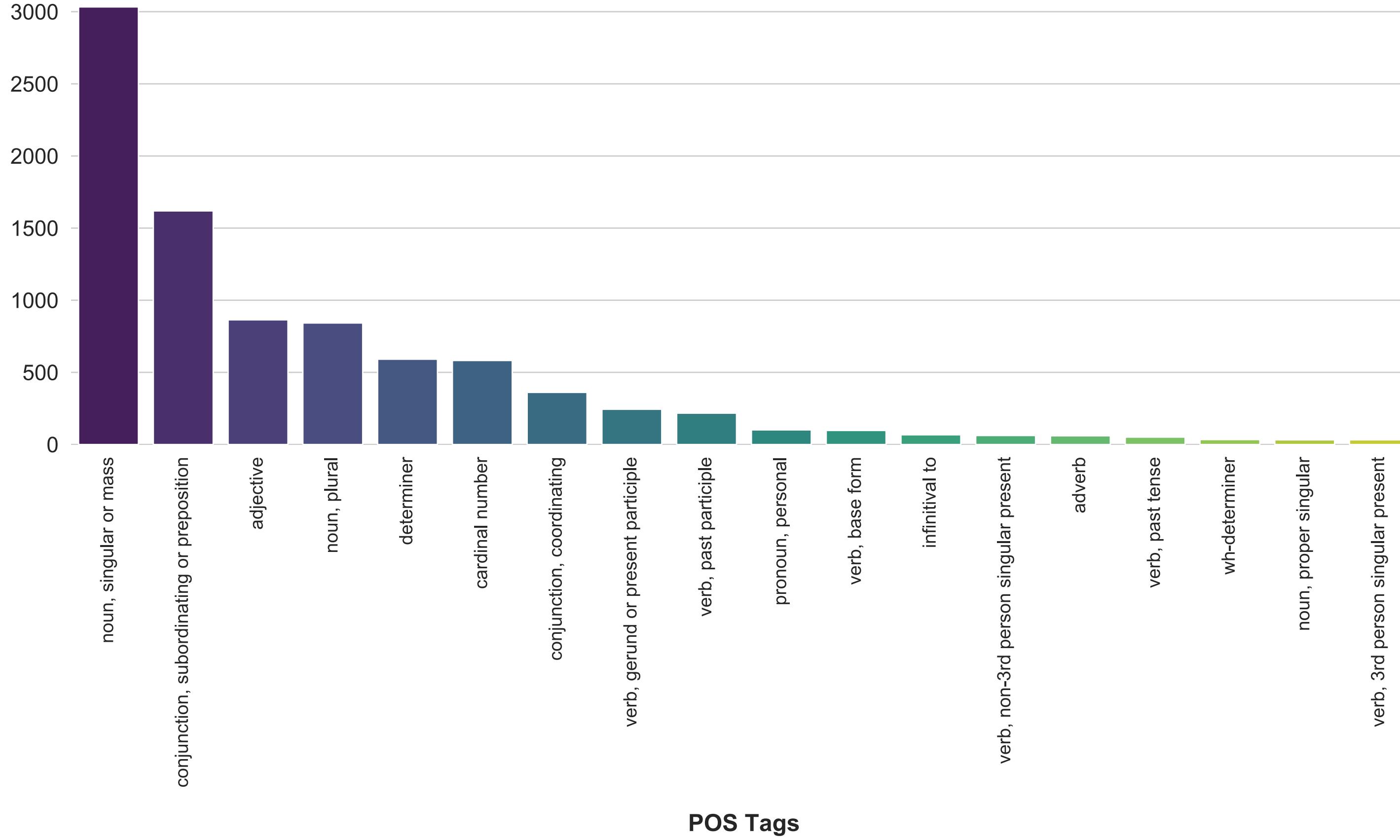


## Full Tokenized Text

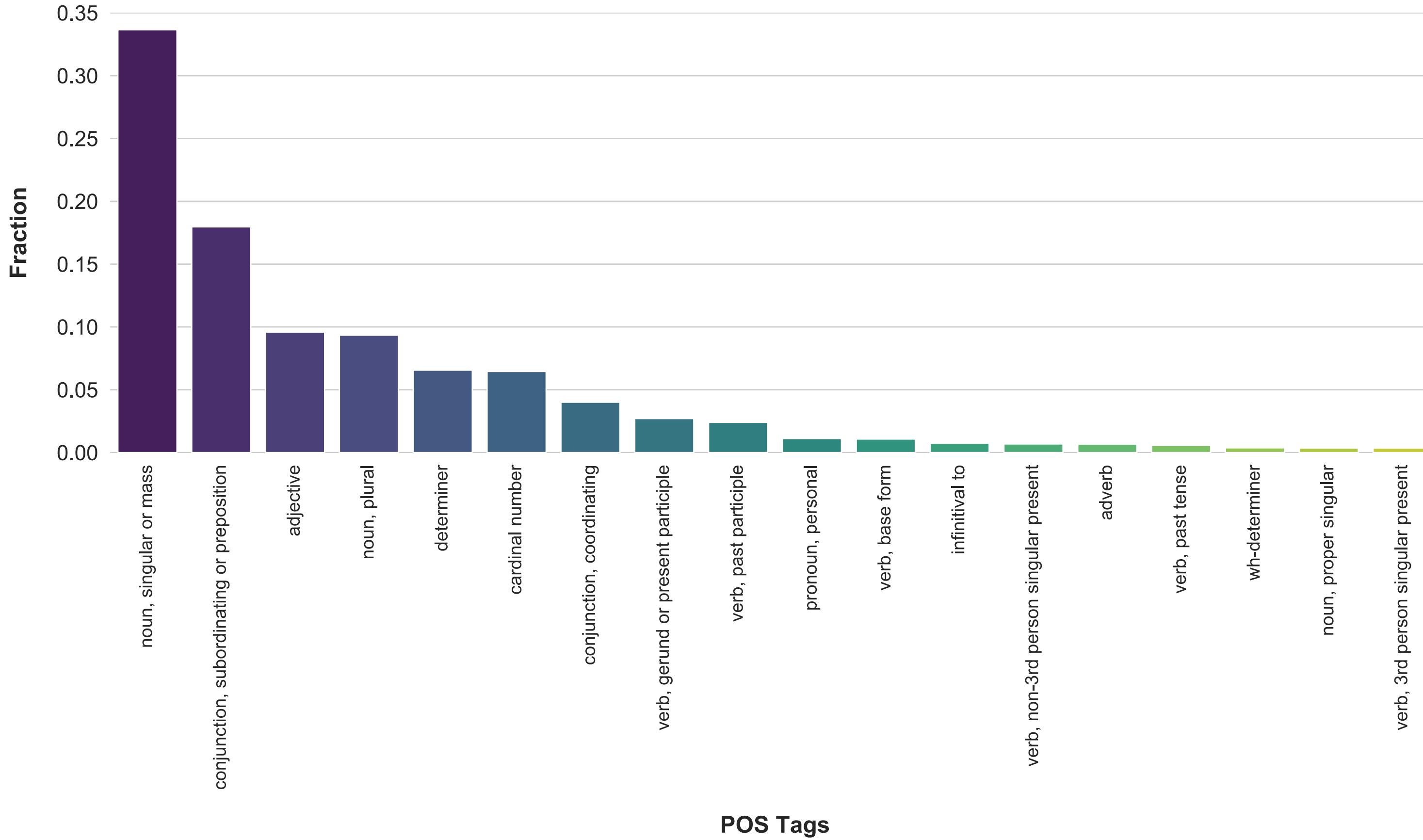


## Full Tokenized Text

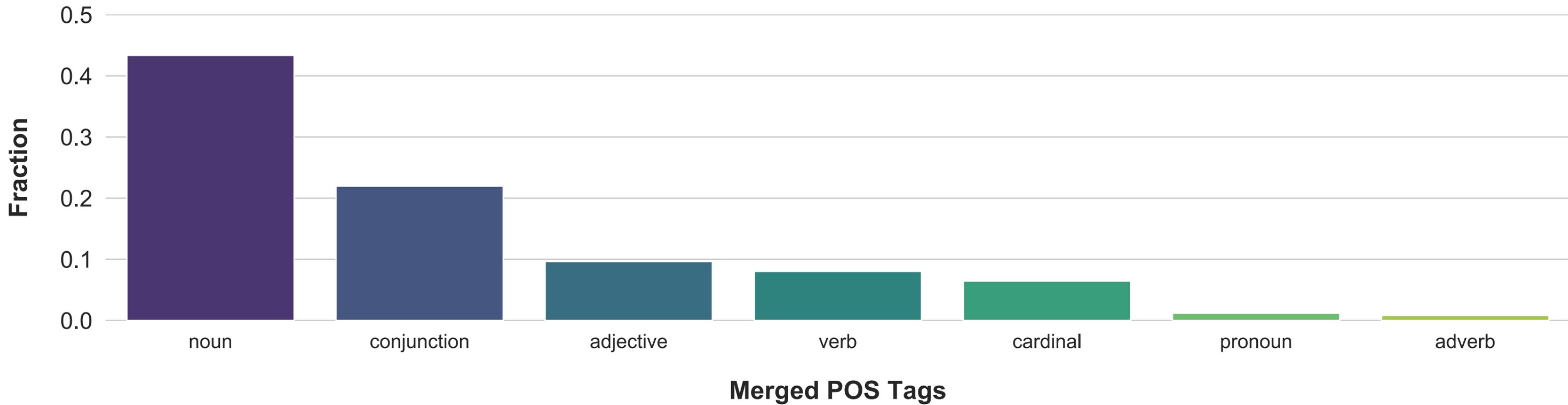
Counts



## Full Tokenized Text



## Full Tokenized Text



## LDA Topic Counts - CountVectorizer

Number of Requests

125  
100  
75  
50  
25  
0

regard, file, record, health  
Topic 1:

waterloo, region, region  
Topic 2:  
waterloo, record

record, regard, relate, file  
Topic 3:

relate, record, waterloo, region  
Topic 4:

environmental, environmental site, assessment, phase  
Topic 5:

include, terminal, guardrail, guardrail, terminal  
Topic 6:

regard, record, waterloo, relate  
Topic 7:

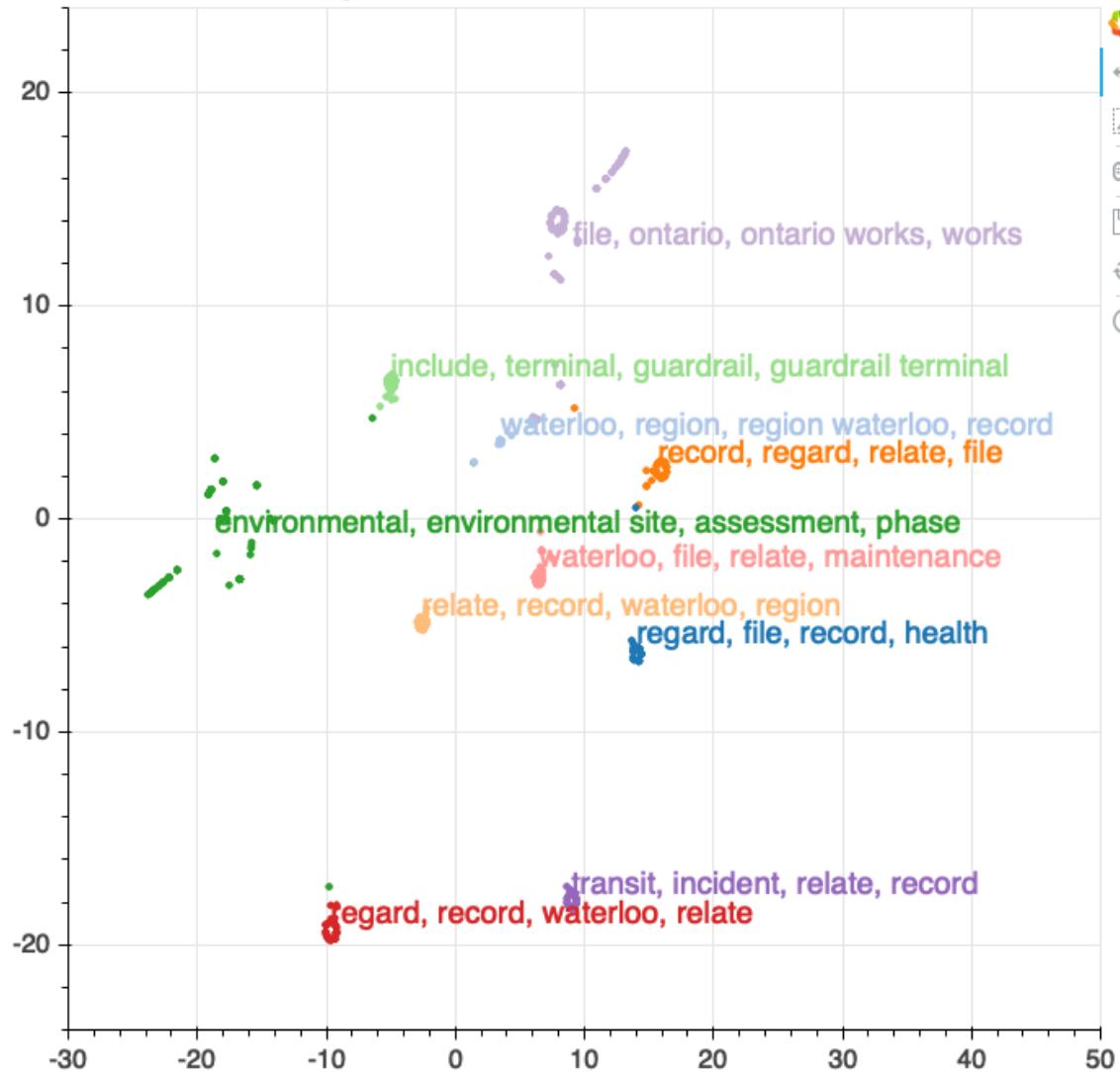
waterloo, file, relate, maintenance  
Topic 8:

transit, incident, relate, record  
Topic 9:

file, ontario, ontario works, works  
Topic 10:

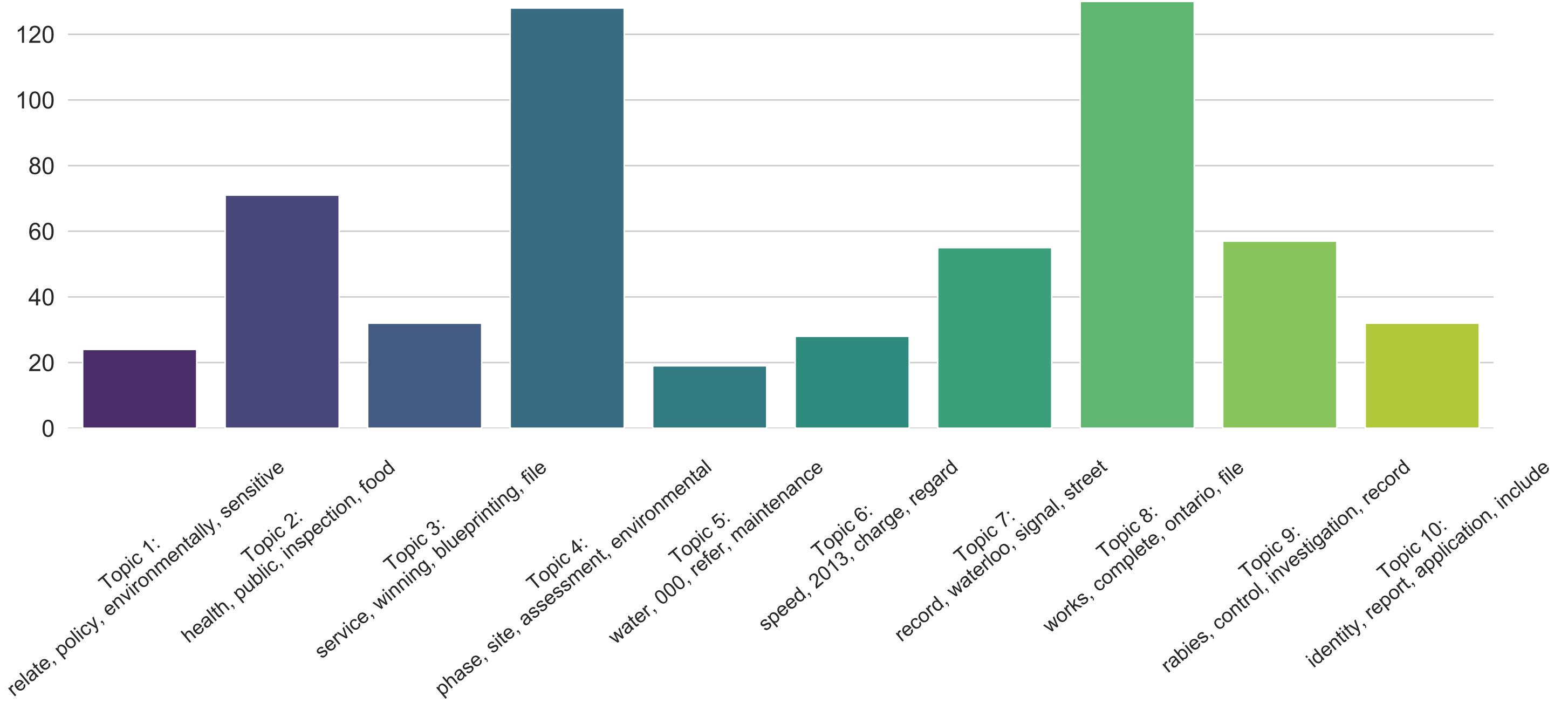


## t-SNE Clustering of 10 LDA Topics - CountVectorizer

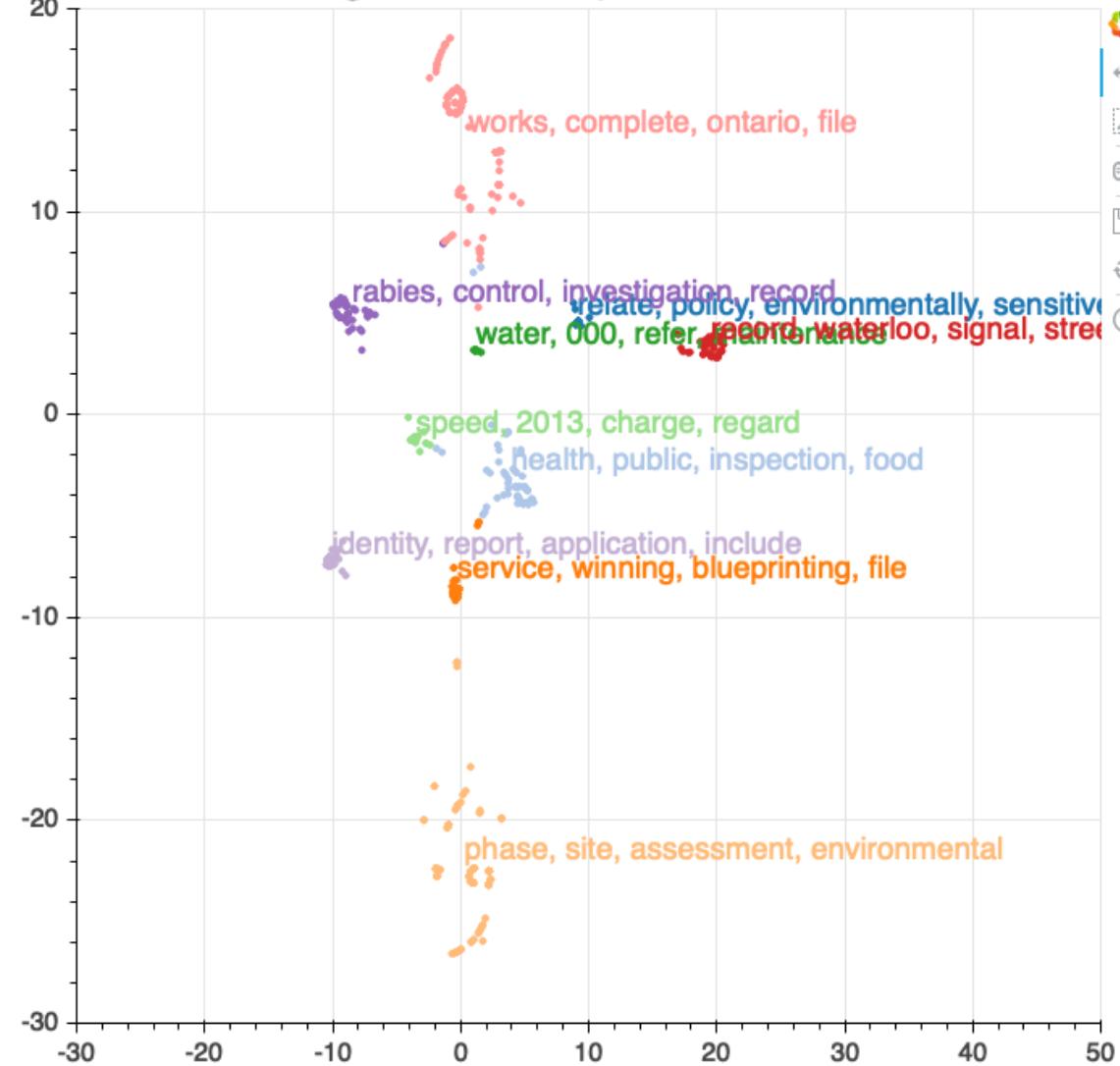


## LDA Topic Counts - tf-idf Vectorizer

Number of Requests



### t-SNE Clustering of 10 LDA Topics - tf-idf Vectorizer



## LSA Topic Counts - CountVectorizer

Number of Requests

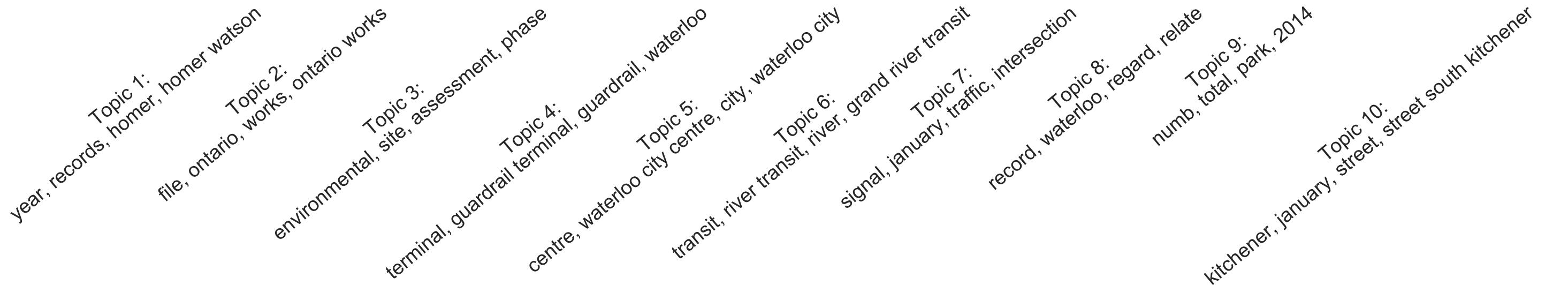
200

150

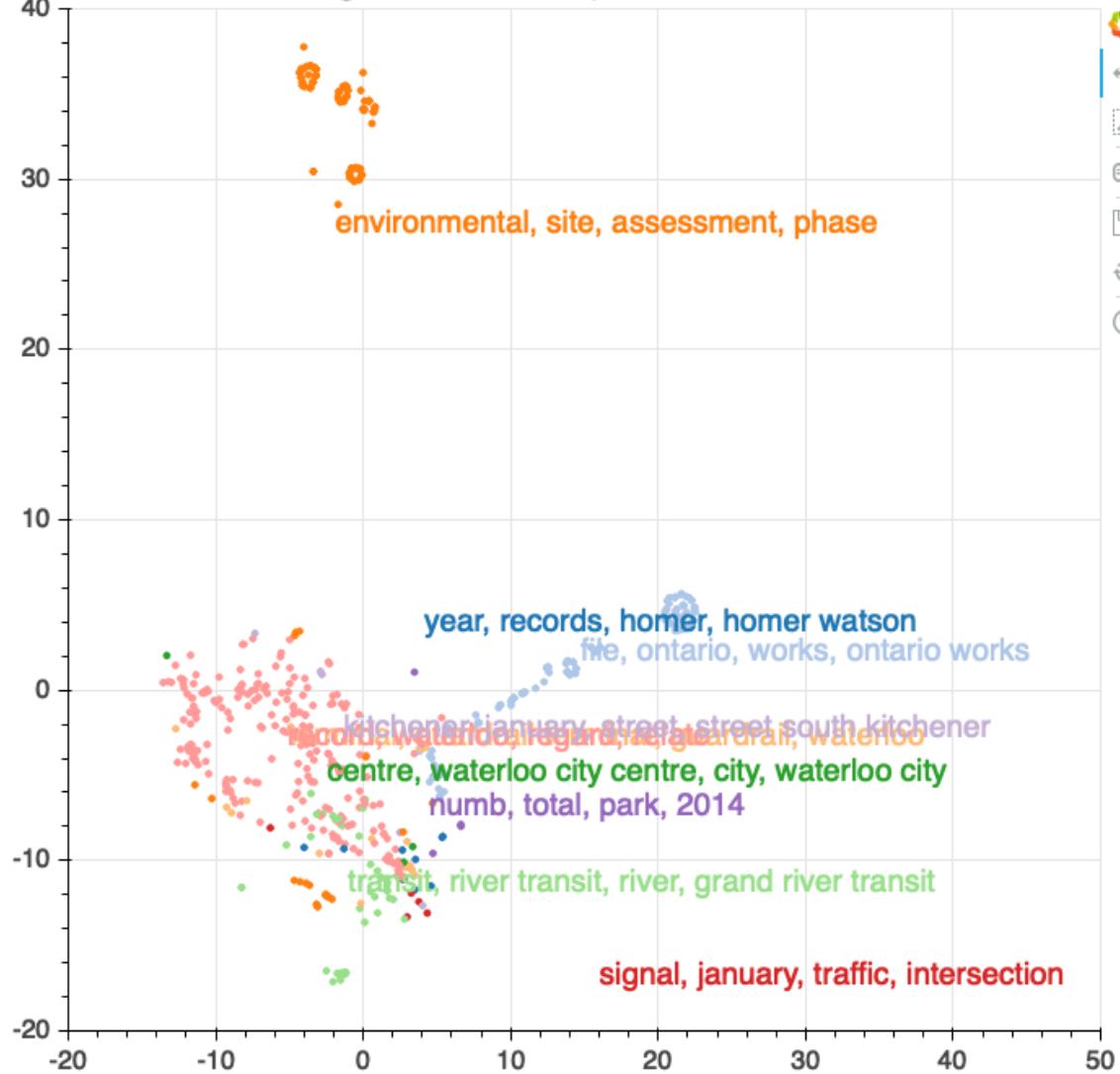
100

50

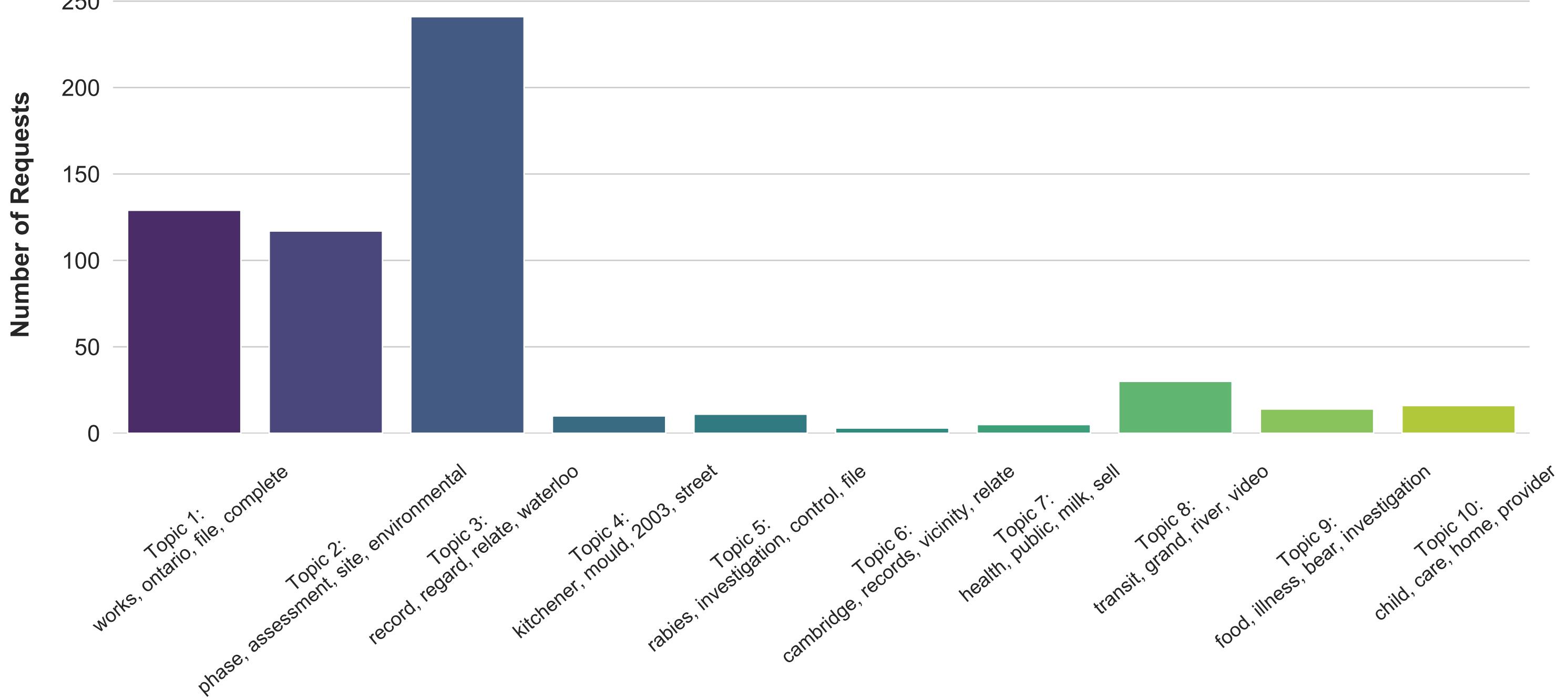
0



## t-SNE Clustering of 10 LSA Topics - CountVectorizer



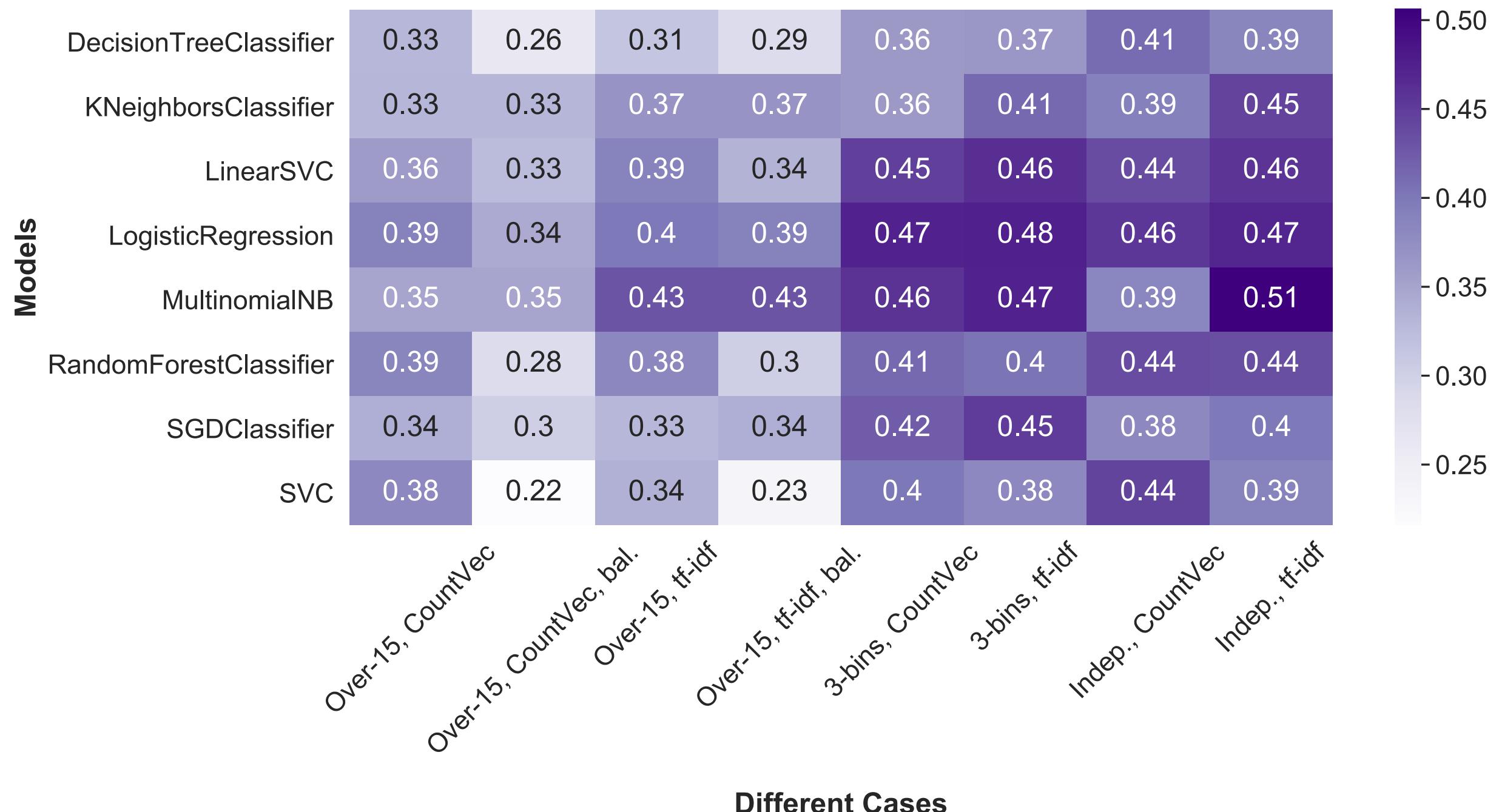
## LSA Topic Counts - tf-idf Vectorizer



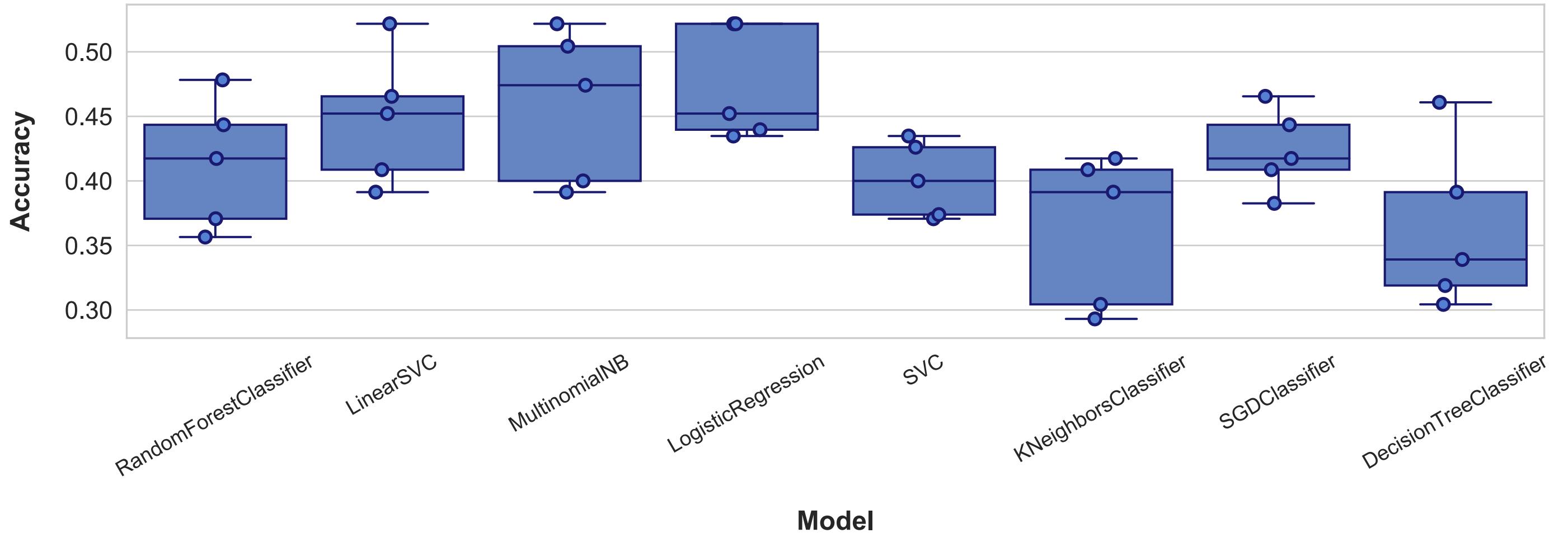
### t-SNE Clustering of 10 LSA Topics - tf-idf Vectorizer



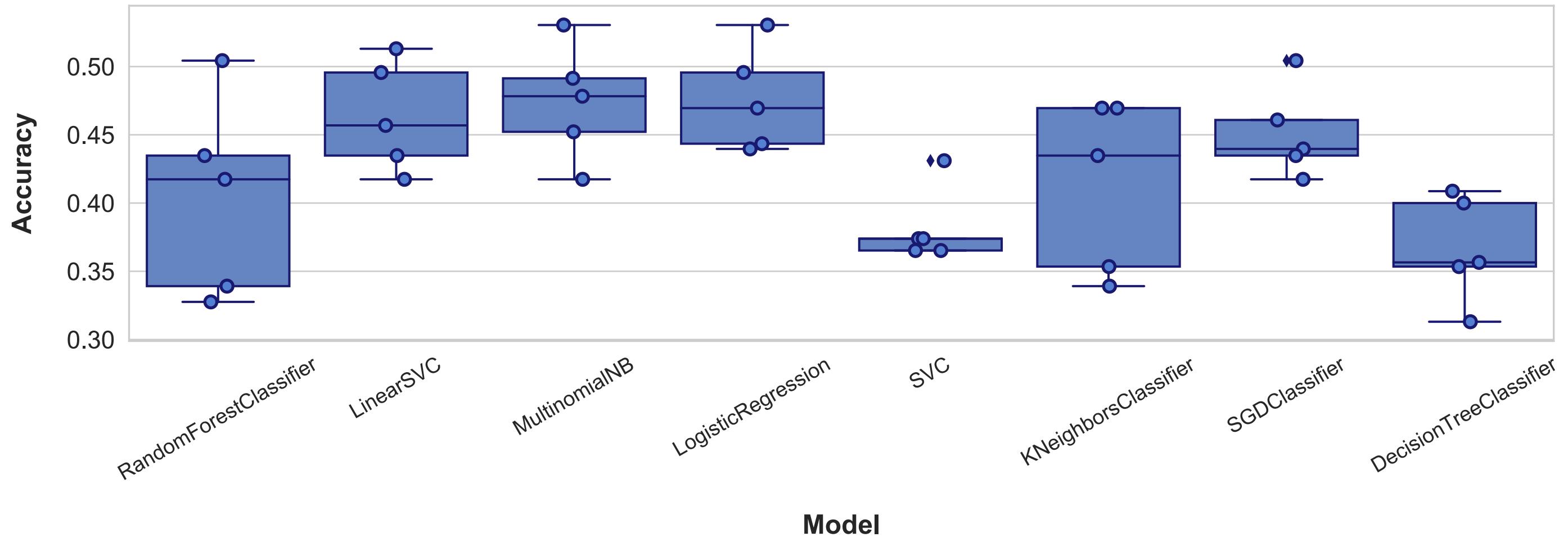
# ML Model Accuracy



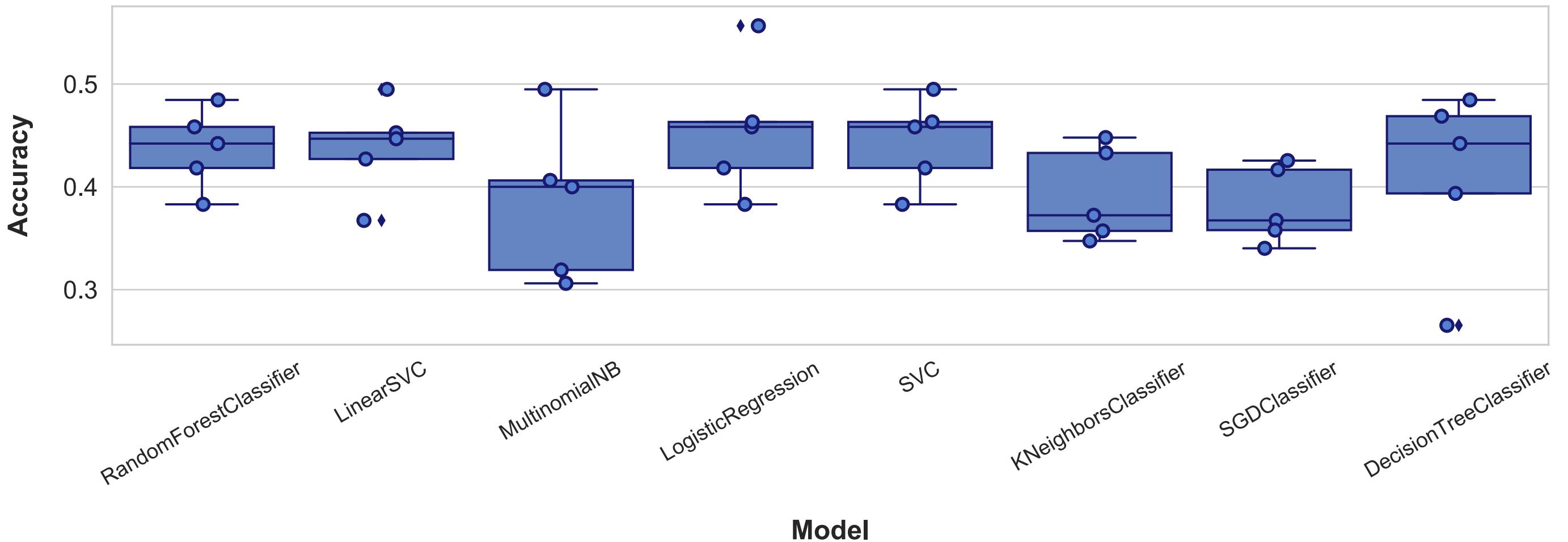
## Classifier comparison for the 3-bin case, using CountVectorizer



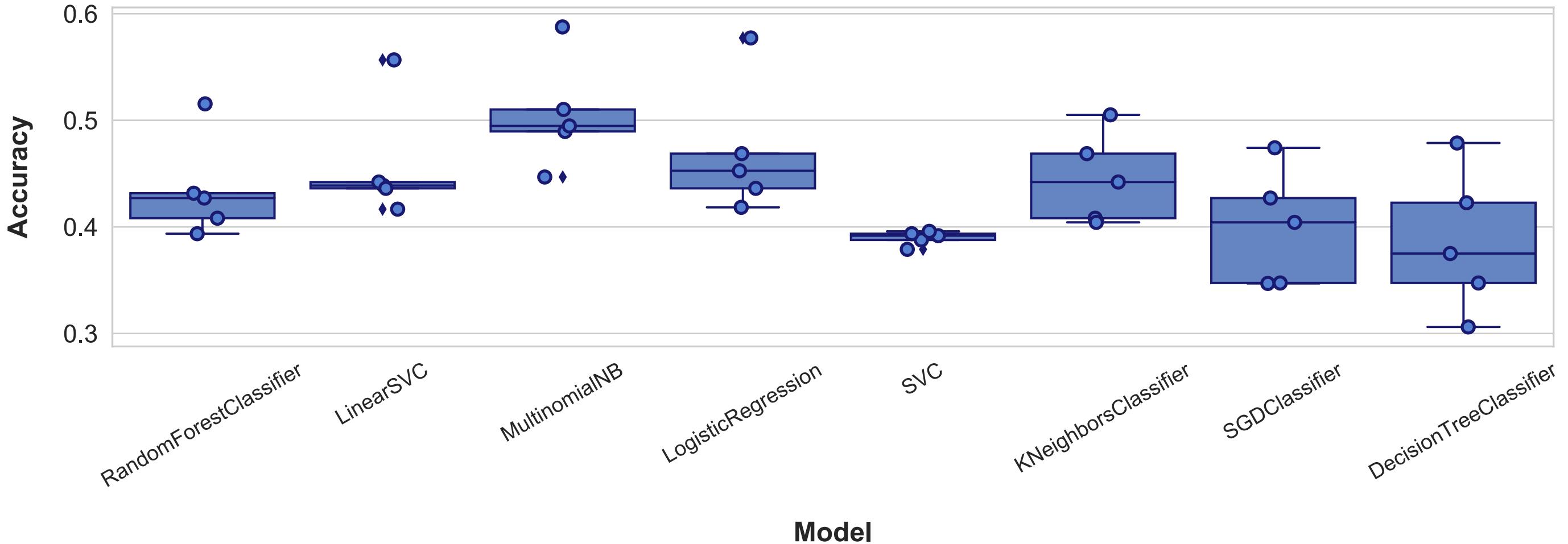
## Classifier comparison for the 3-bin case, using tf-idf



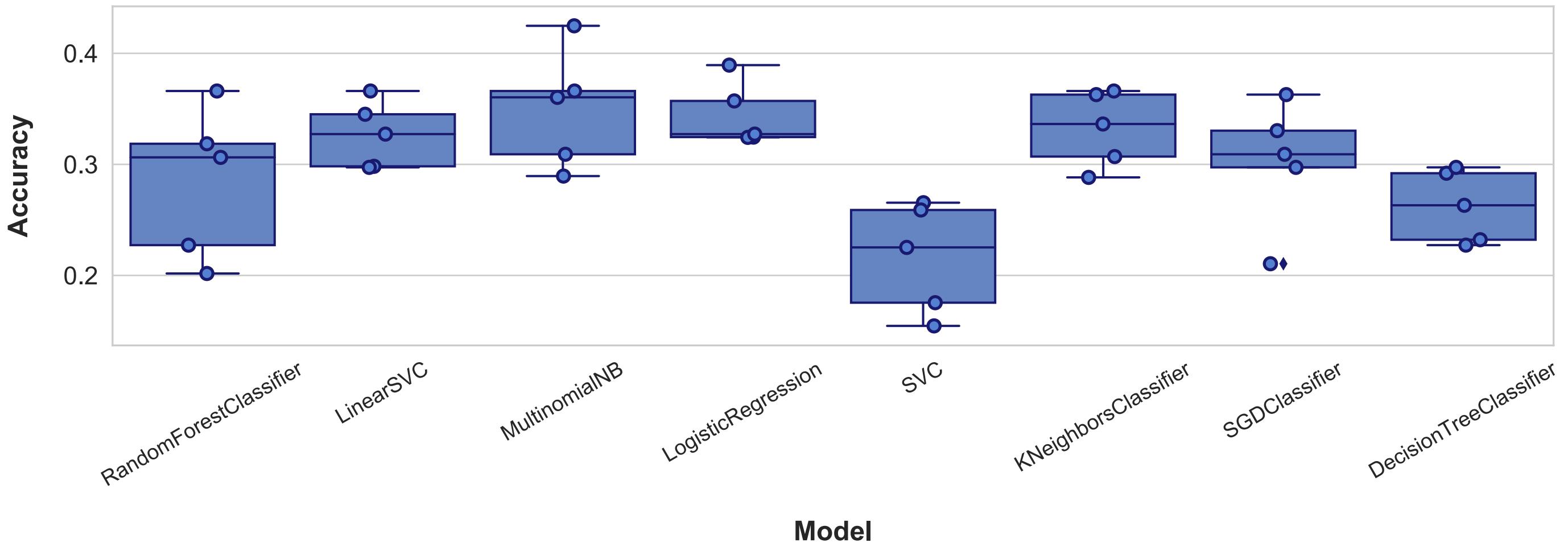
## Classifier comparison for the indep. case, using CountVectorizer



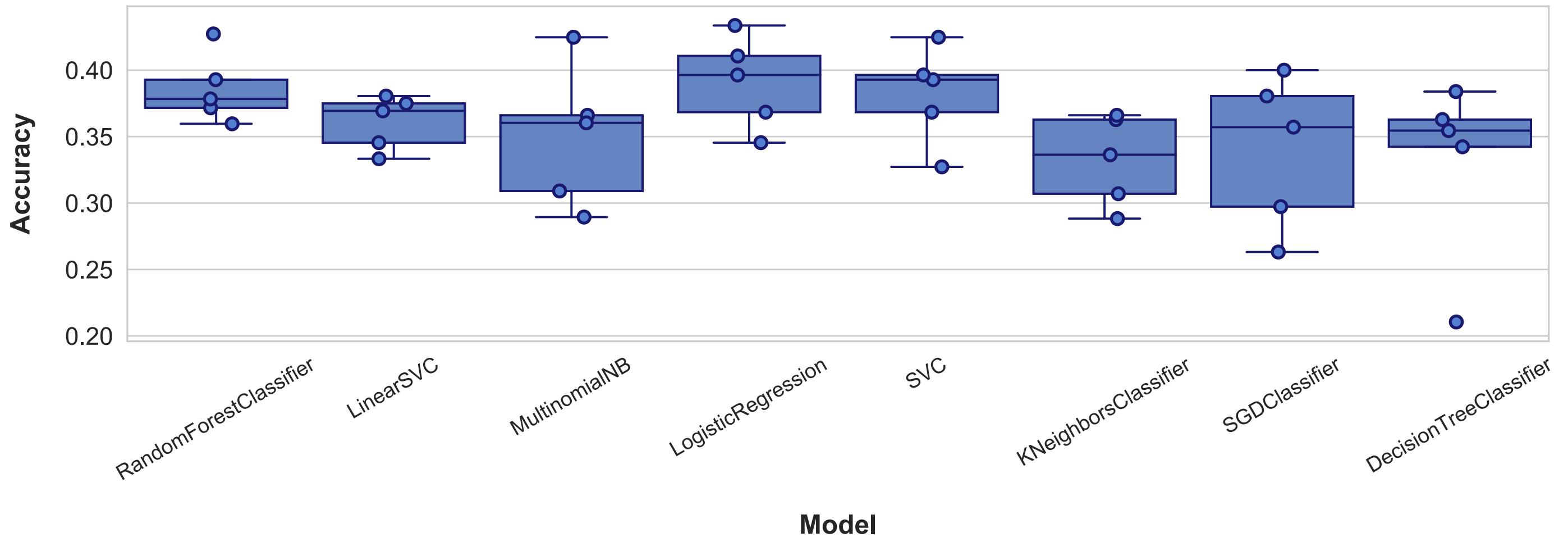
## Classifier comparison for the indep. case, using tf-idf



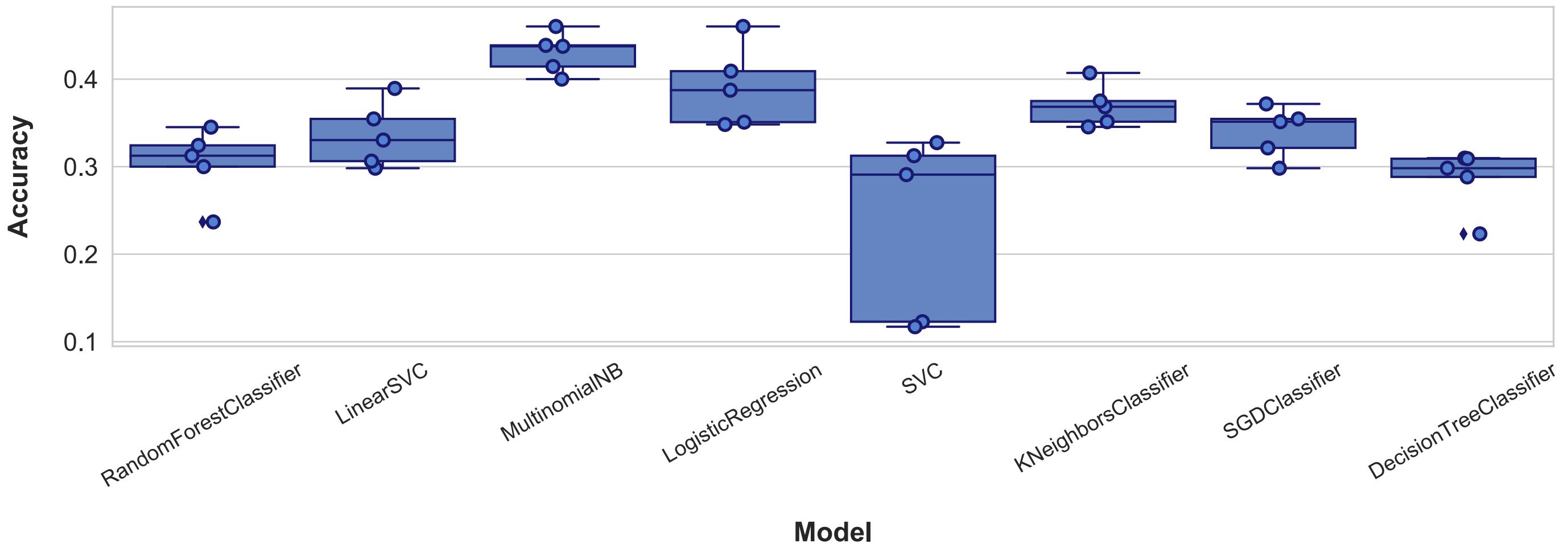
## Classifier comparison for the over 15 case, using CountVectorizer, Balanced



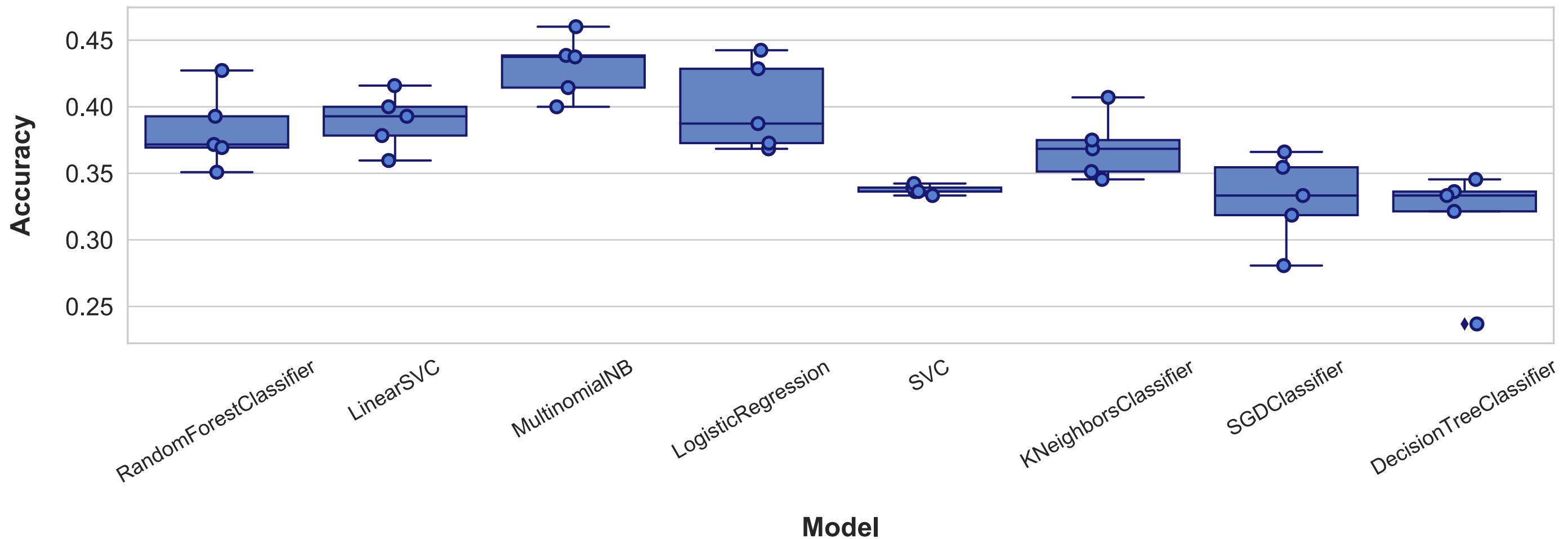
## Classifier comparison for the over 15 case, using CountVectorizer



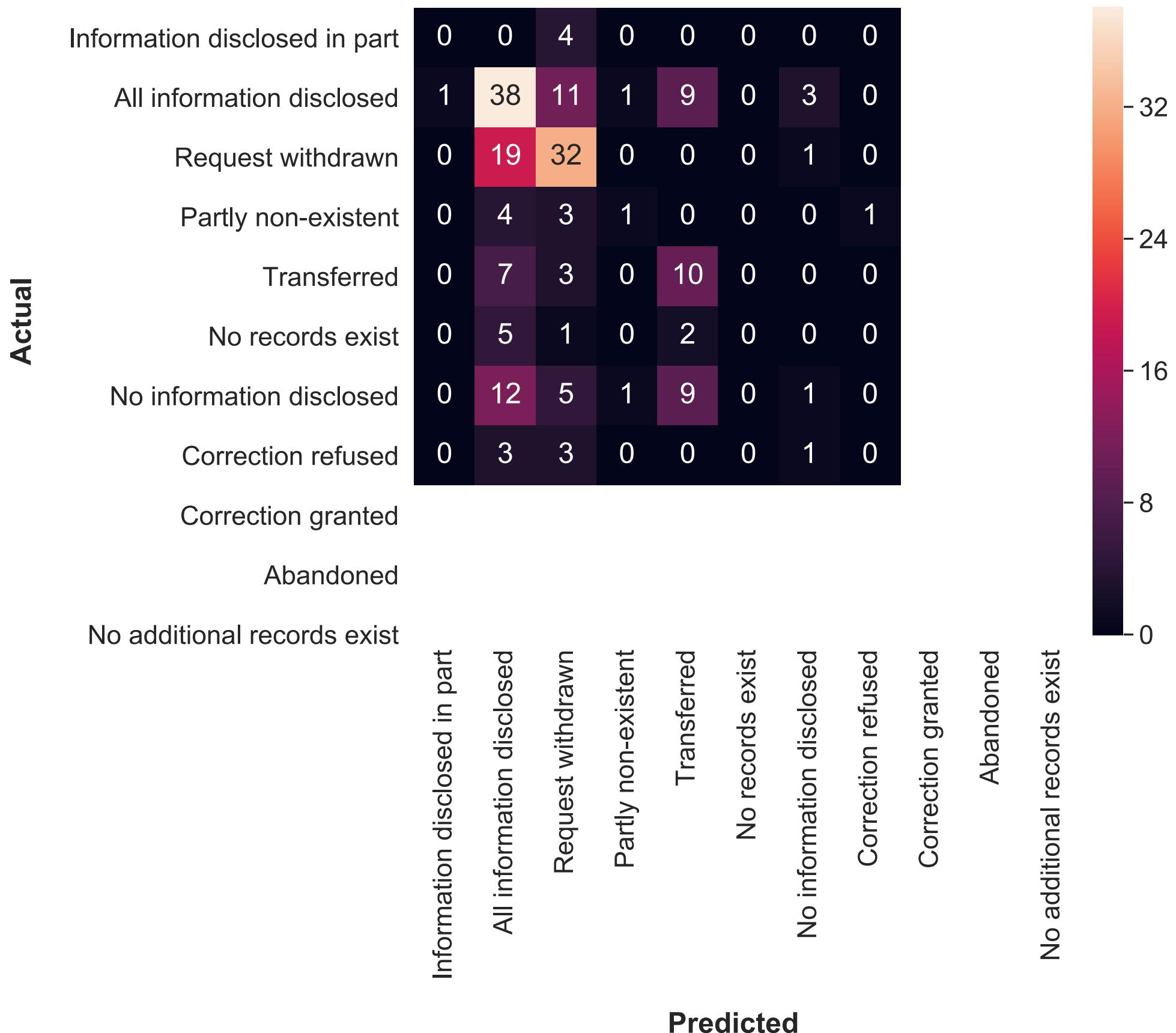
## Classifier comparison for the over 15 case, using tf-idf, balanced



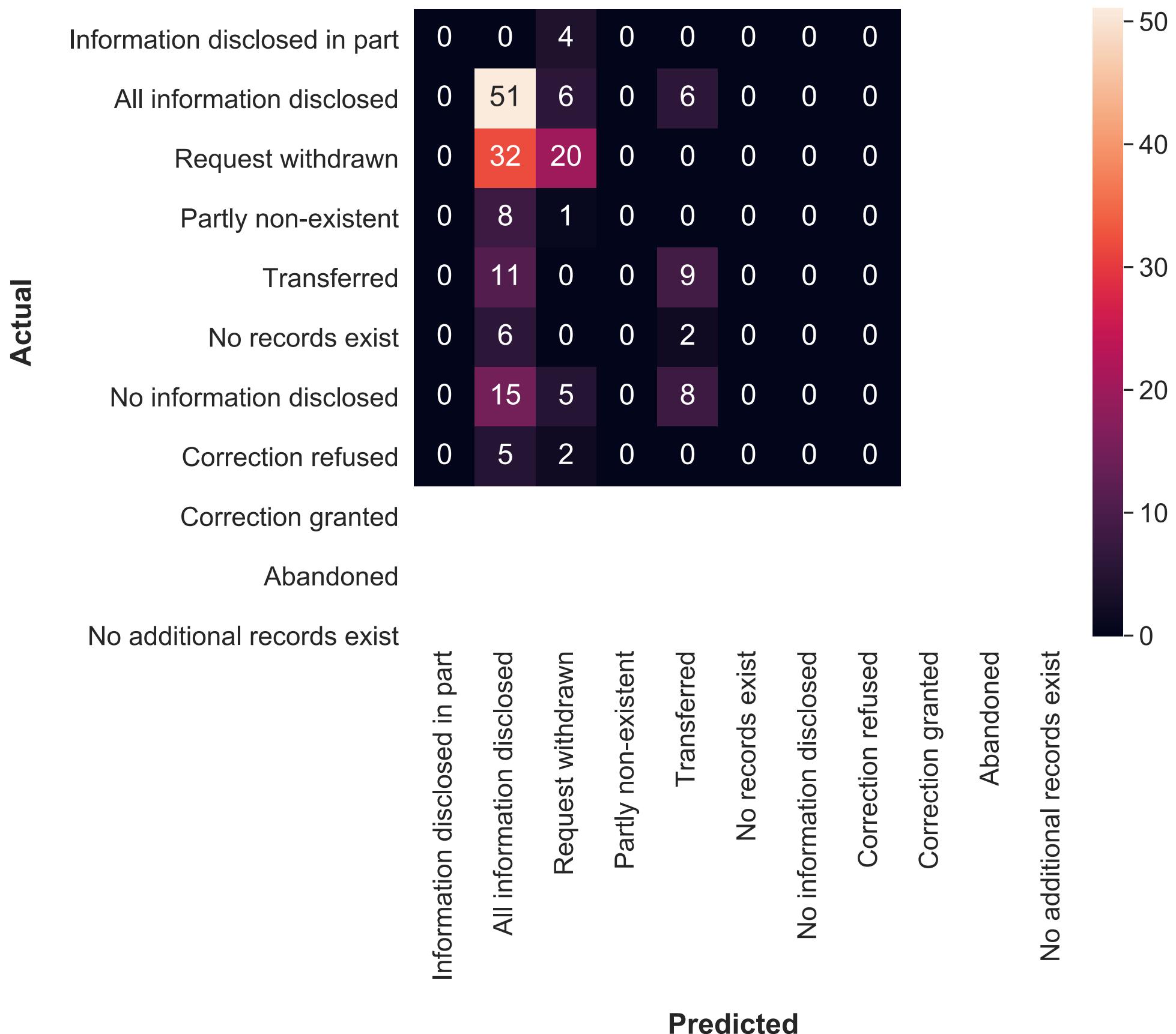
## Classifier comparison for the over 15 case, using tf-idf



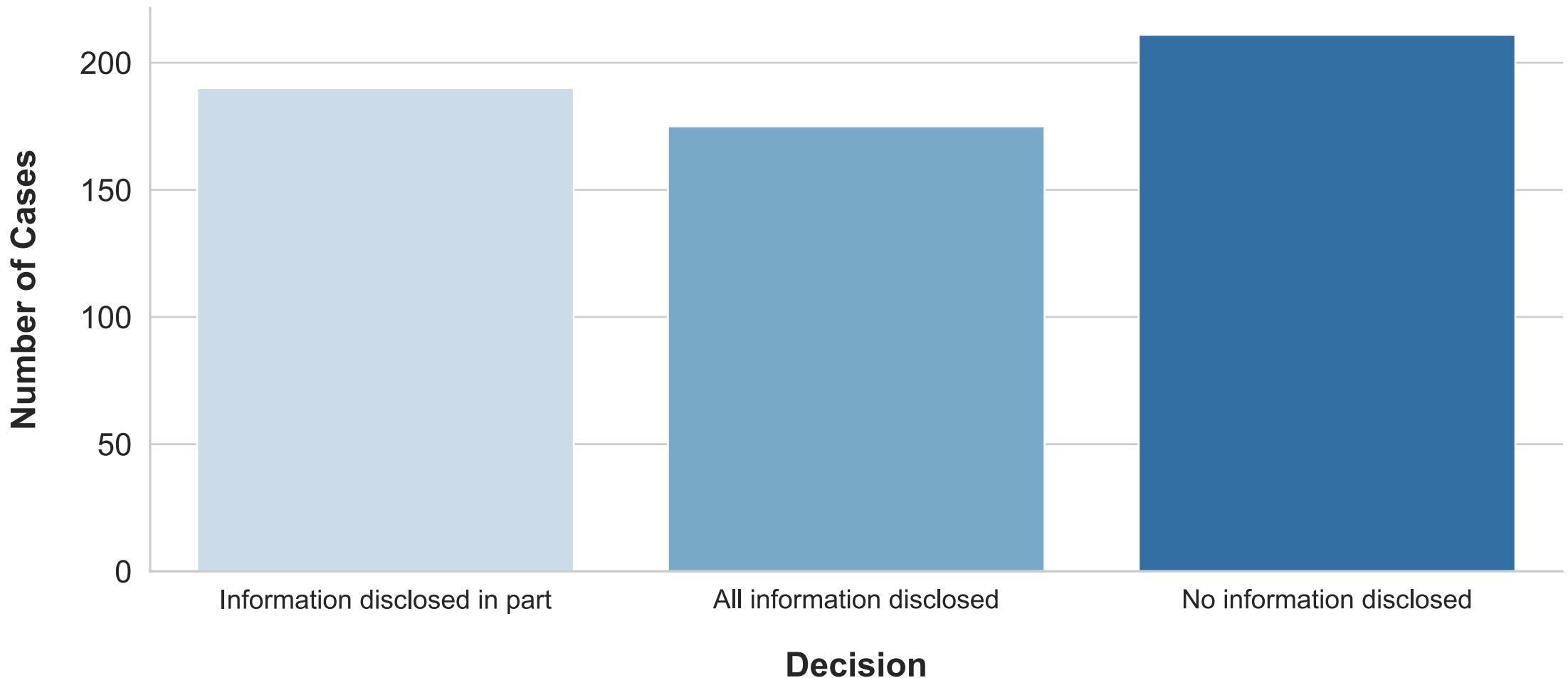
# MultinomialNB, CountVectorizer, full set



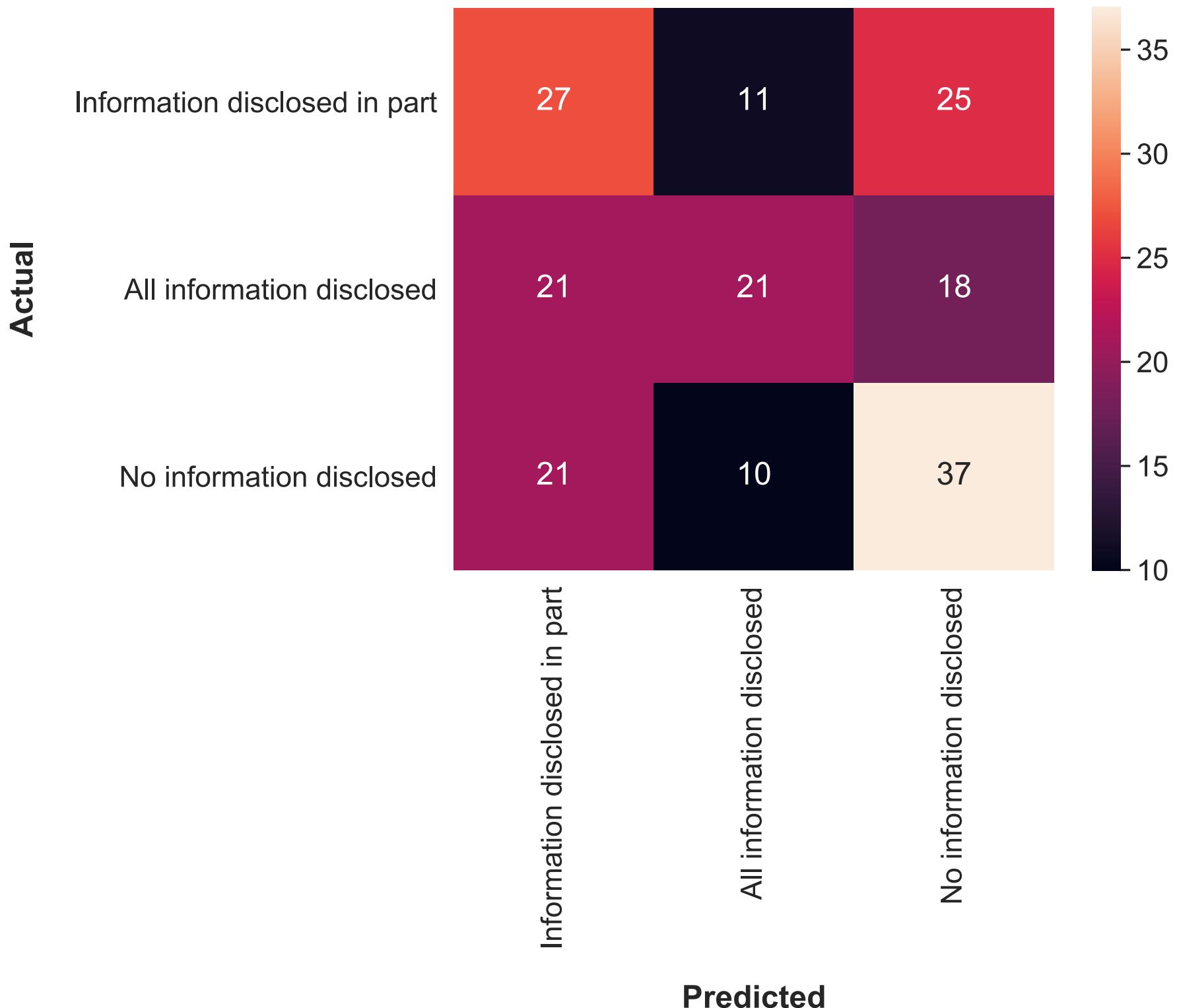
## MultinomialNB, tf-idf, full set



## **Full data split into three categories only**

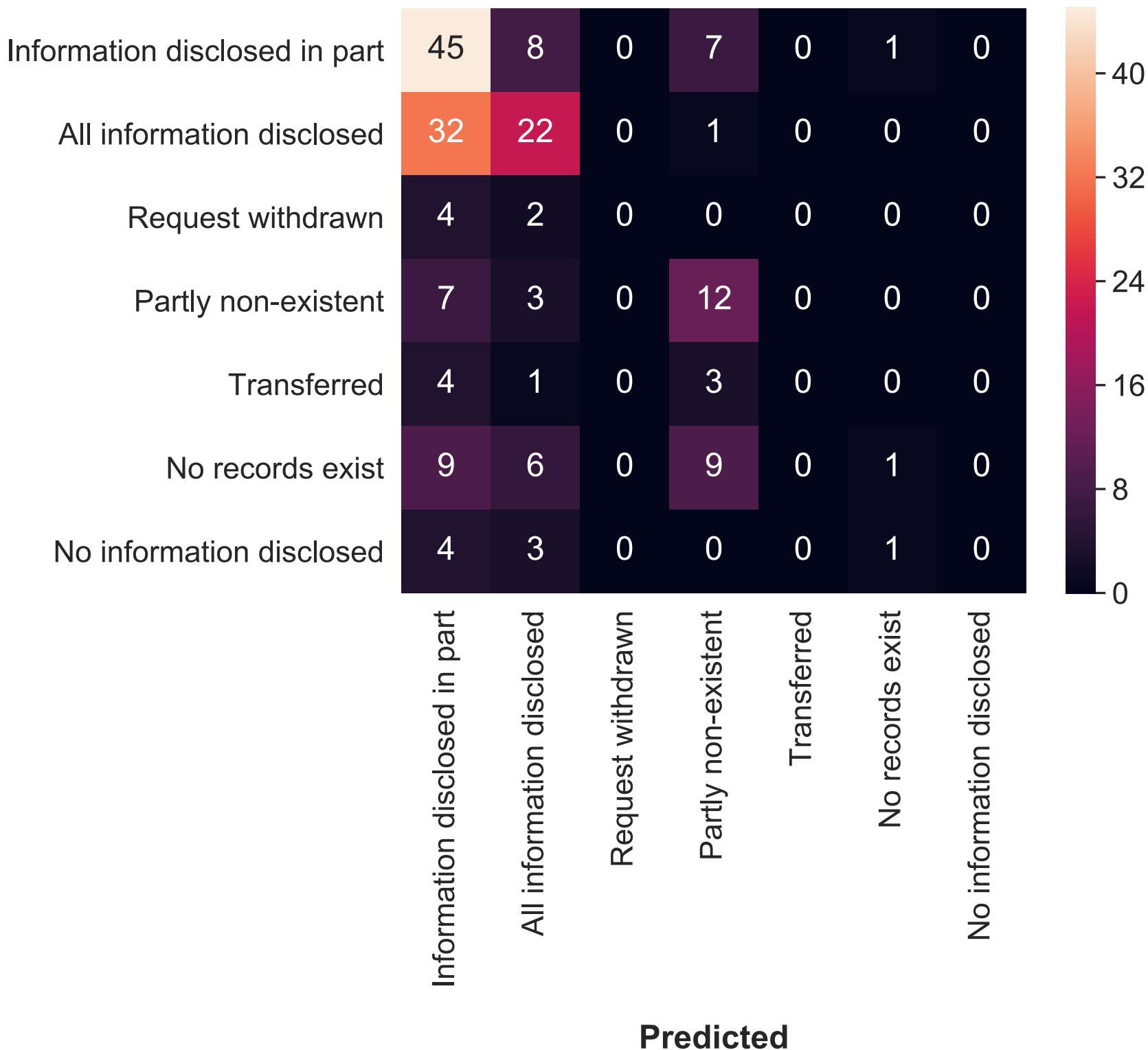


## LogisticRegression, tf-idf, 3 bins



## MultinomialNB, tf-idf, over 15

Actual



## MultinomialNB, tf-idf, indep.

