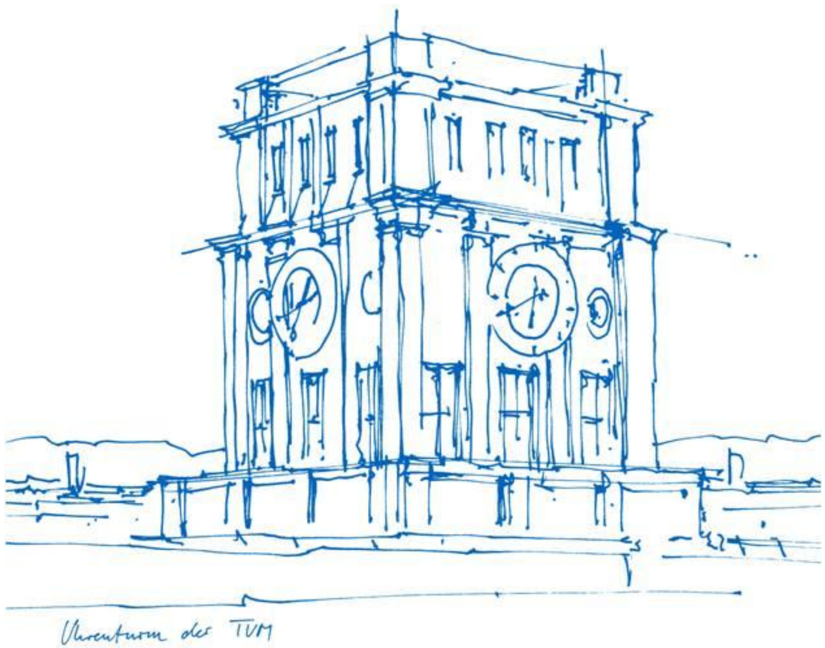
Understanding Class(ifier) Differences

XAI Lab Course

WS 20/21
February 9, 2021

Technical University of Munich
Faculty of Informatics



Kadir Aslan, Felix Dietrich

Explainability for
Understanding Class(ifer) Differences

Introduction

- Understand **class** differences and **classifier** differences
  - Classification Task: **Hate Speech Detection (NLP)**
  - Analyze multiple datasets
  - Compare multiple classification methods
    - By performance metrics
    - By applying explainability methods
- Attempted systematic approach

Hate Speech Datasets

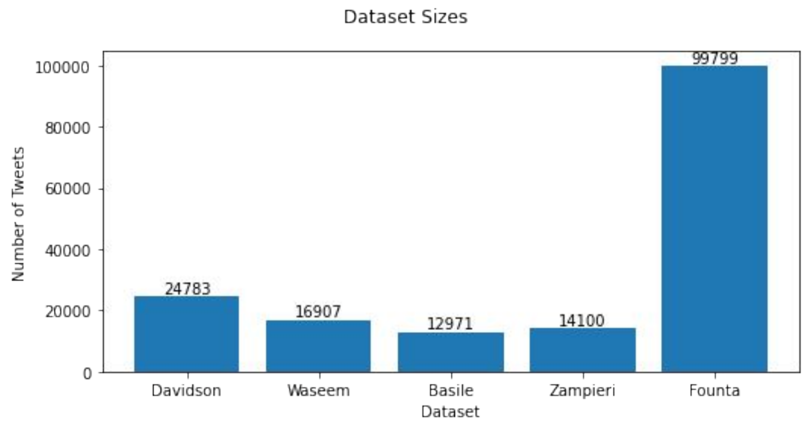
- Lots of datasets[1] from different sources available
- Focus on **English**, **Twitter**, and **text only**

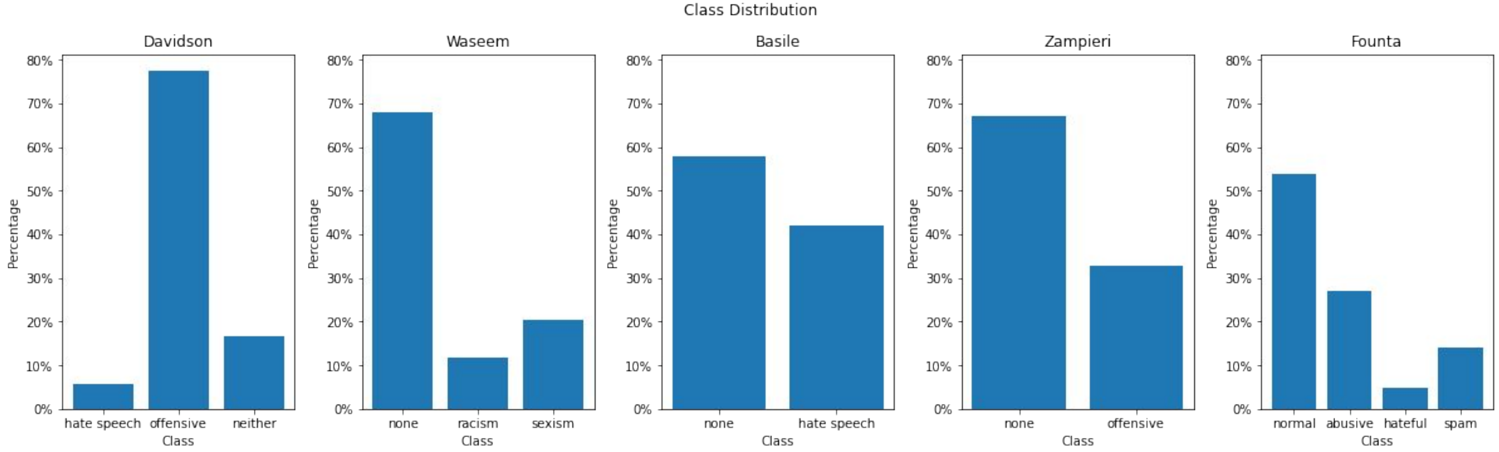
Dataset id	# Instances	Classes
Waseem (2016)	16907	racism, sexism, none
Zampieri (2019)	14100	offensive, none
Founta (2018)	99799	abusive, hateful, spam, normal
Basile (2019)	12971	hate-speech, none
Davidson (2017)	24783	hate-speech, offensive, neither

[1]https://hatespeechdata.com

Datasets Overview

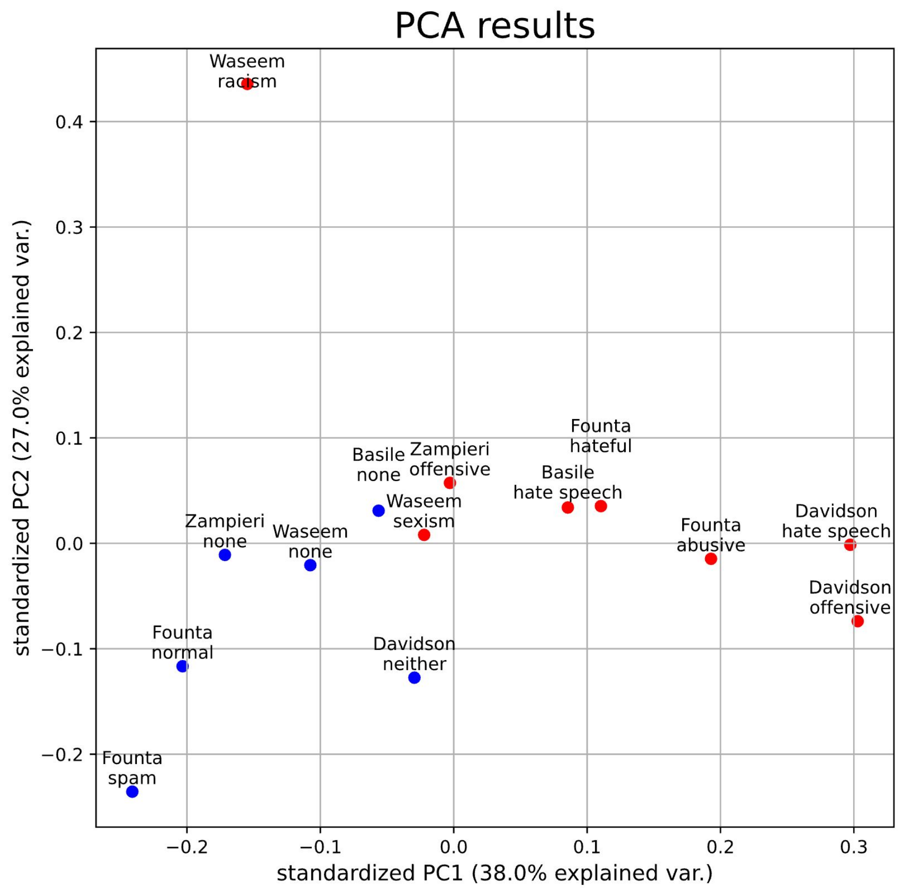
- Different kinds of abuse
- Imbalances
- Different data collection strategies





Inter-Dataset Class Similarity

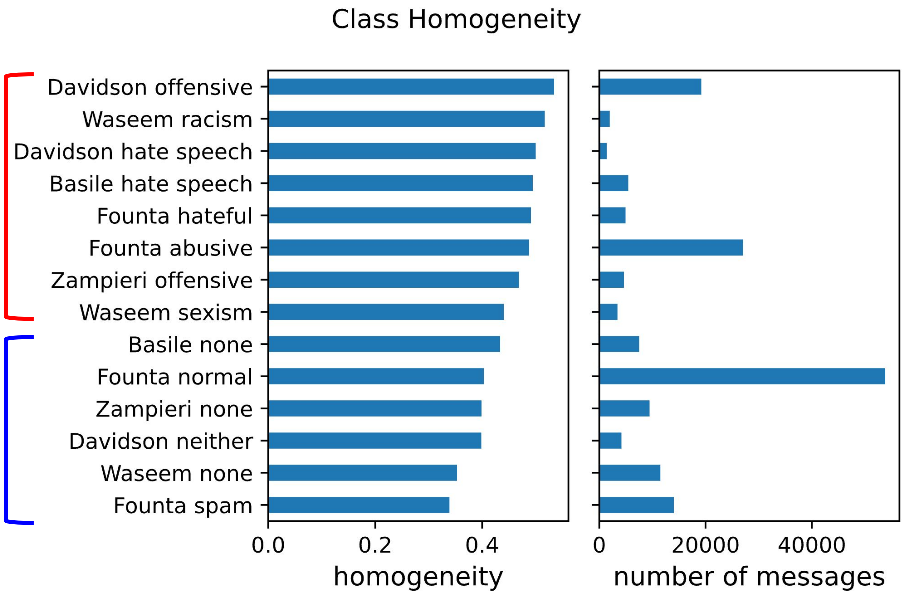
1. Preprocess
2. FastText pre-trained embeddings
3. Calculate tweet centroids
4. Group by classes
5. Average → class centroids
6. PCA



Fortuna et al.: "Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets"

Intra-Category Homogeneity

1. Preprocess
2. FastText pre-trained embeddings
3. Calculate tweet centroids
4. Group by classes
5. Calculate cosine similarity matrix
6. Average entries



Fortuna et al.: "**Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets**"

(Macro) F1 Scores

- Viable classification options in **scikit**
- Simple neural models with **TensorFlow**

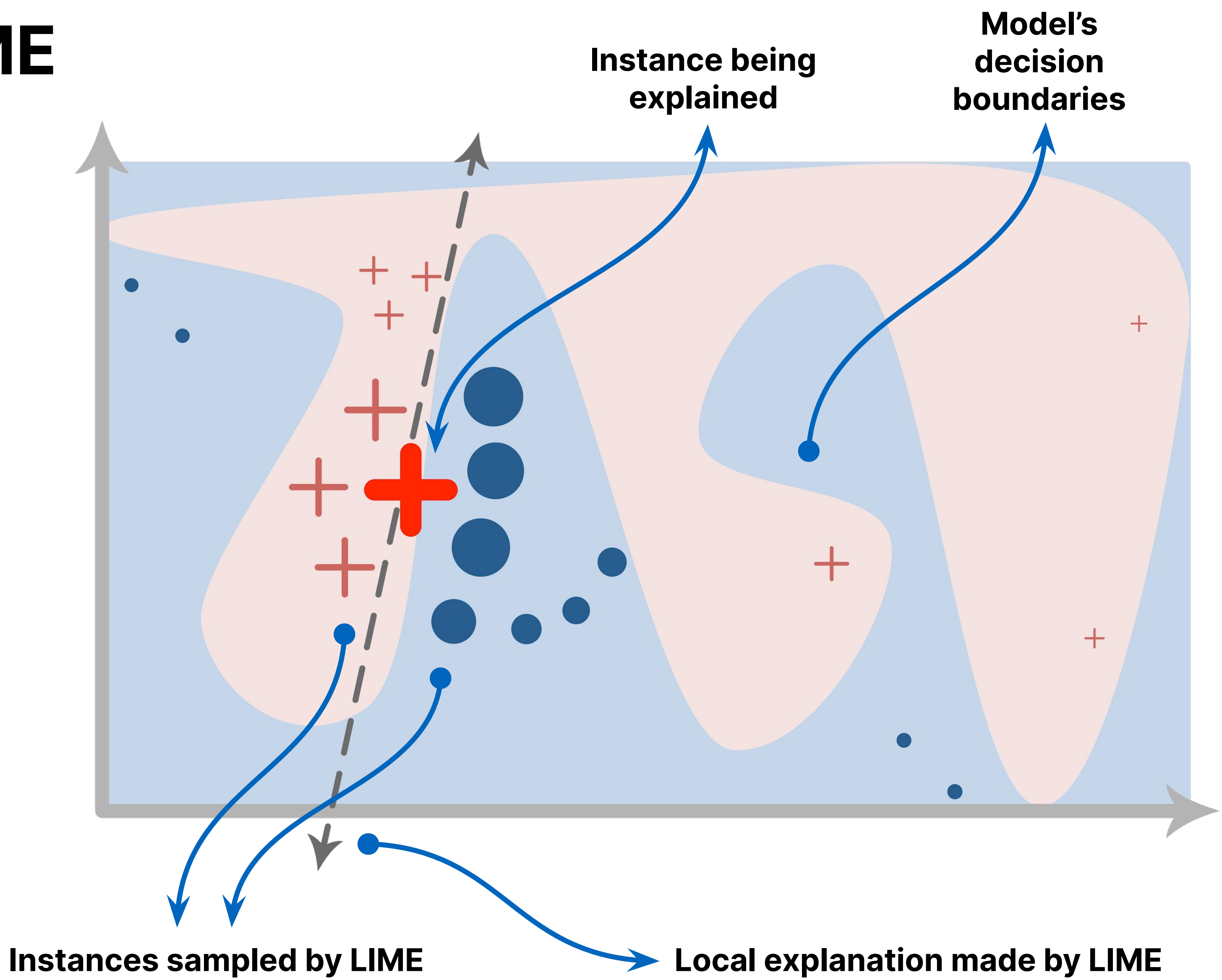
Binarized union of all datasets

		Davidson	Waseem	Basile	Zampieri	Founta	Combined
scikit	LinearSVC	68.7	76.9	72.5	70.5	64.2	87.8
	GaussianNB	53.6	41.4	64.0	55.6	45.7	72.7
	ComplementNB	61.0	73.2	73.2	63.0	55.9	83.3
	DecisionTreeClassifier	66.6	71.8	66.7	64.3	59.2	85.0
	KNeighborsClassifier	55.2	65.4	66.6	60.8	48.2	73.8
	RandomForestClassifier	56.6	73.7	71.7	66.4	57.1	86.9
	MLPClassifier	68.3	72.9	69.8	66.9	61.6	85.1
tf	DenseClassifier	67.1	74.3	70.7	68.0	63.2	86.6
	LSTMClassifier	61.6	73.0	70.0	66.9	64.2	87.9
	CNNClassifier	62.7	44.1	70.5	69.4	64.1	87.8

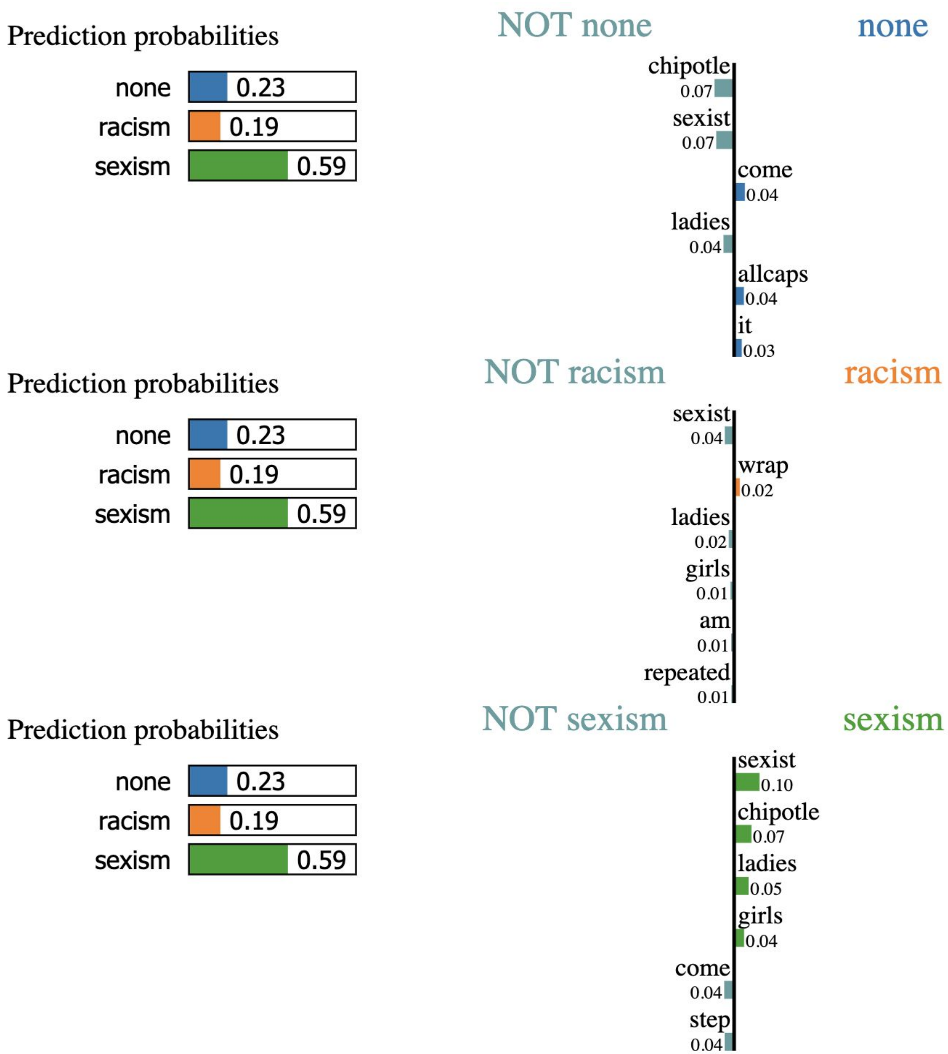
LIME

- **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations
- Computes **feature importance** scores
- Black-box model's decision function is approximated with a **locally** faithful model.
    - LIME samples instances
    - Gets predictions using the original model
    - Weights them by their distance to the instance being explained

LIME


Instance being explained
Model's decision boundaries
Instances sampled by LIME
Local explanation made by LIME

Local Explanation - Complement NB



Text with highlighted words

rt |userl : i am so not sexist but girls cannot wrap burritos at chipotle . |repeated| i think it ' s known fact . |allcaps| come on ladies step it up |/allcaps|

Text with highlighted words

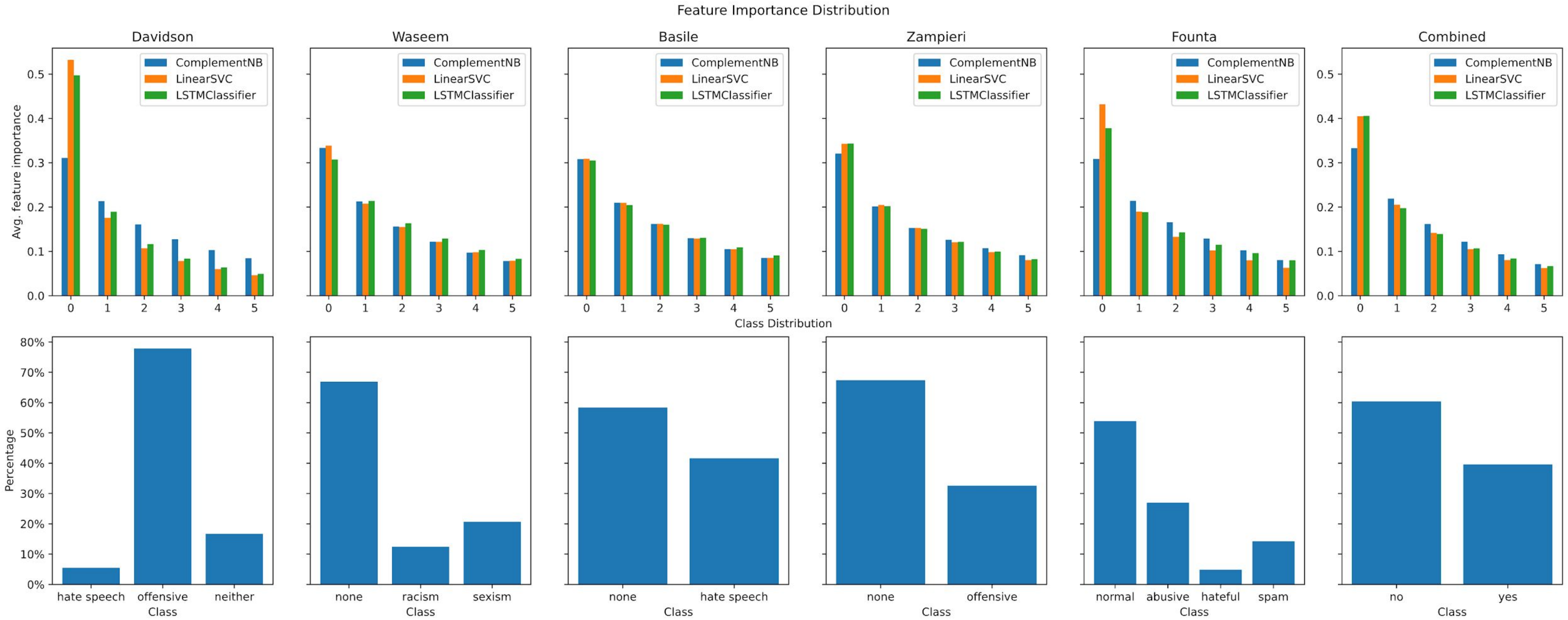
rt |userl : i am so not sexist but girls cannot wrap burritos at chipotle . |repeated| i think it ' s known fact . |allcaps| come on ladies step it up |/allcaps|

Text with highlighted words

rt |userl : i am so not sexist but girls cannot wrap burritos at chipotle . |repeated| i think it ' s known fact . |allcaps| come on ladies step it up |/allcaps|

Feature Importance Distributions

Absolute value → sort → normalize → average → normalize



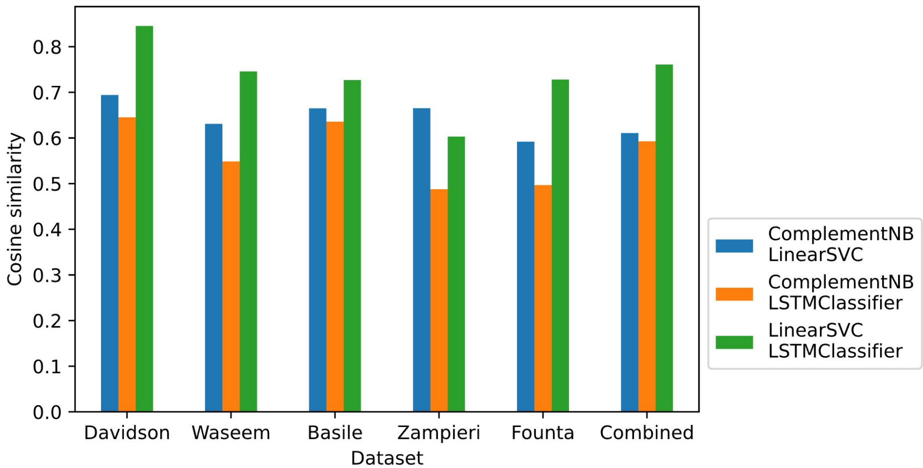
Explanations over 3500 instances with ≥ 6 features

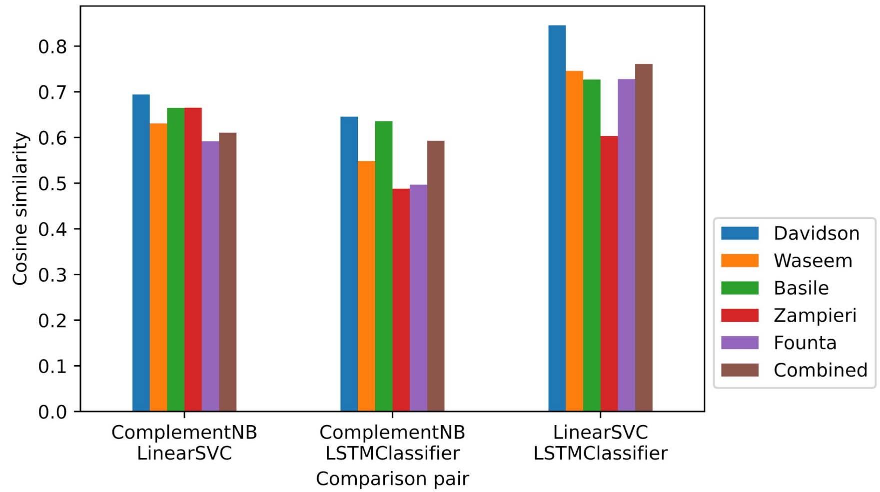
F1 scores

	Davidson	Waseem	Basile	Zampieri	Founta	Combined
ComplementNB	61.0	73.2	73.2	63.0	55.9	83.3
LinearSVC	68.7	76.9	72.5	70.5	64.2	87.8
LSTMClassifier	61.6	73.0	70.0	66.9	64.2	87.9

Feature Importance Similarities

- **Generative** classifier: ComplementNB
- **Discriminative** classifier: LinearSVC, LSTMClassifier





F1 scores

	Davidson	Waseem	Basile	Zampieri	Founta	Combined
ComplementNB	61.0	73.2	73.2	63.0	55.9	83.3
LinearSVC	68.7	76.9	72.5	70.5	64.2	87.8
LSTMClassifier	61.6	73.0	70.0	66.9	64.2	87.9

Classifier Prediction Stability

Observe prediction changes by omitting the most important feature word

Davidson

	F1	F1 (omitted)	abs.	rel. %
ComplementNB	57.27	45.30	-11.97	-20.90
LinearSVC	66.33	44.93	-21.39	-32.26
LSTMClassifier	60.16	39.85	-20.31	-33.76

Waseem

	F1	F1 (omitted)	abs.	rel. %
ComplementNB	73.80	60.33	-13.47	-18.26
LinearSVC	76.95	56.05	-20.89	-27.15
LSTMClassifier	72.93	54.60	-18.33	-25.13

Basile

	F1	F1 (omitted)	abs.	rel. %
ComplementNB	68.82	53.03	-15.79	-22.94
LinearSVC	67.12	48.63	-18.49	-27.55
LSTMClassifier	65.46	46.58	-18.88	-28.85

Zampieri

	F1	F1 (omitted)	abs.	rel. %
ComplementNB	40.77	13.62	-27.15	-66.58
LinearSVC	55.92	22.33	-33.59	-60.07
LSTMClassifier	53.52	25.71	-27.81	-51.95

Founta

	F1	F1 (omitted)	abs.	rel. %
ComplementNB	55.57	46.14	-9.43	-16.96
LinearSVC	61.92	34.09	-27.83	-44.95
LSTMClassifier	61.95	36.44	-25.51	-41.17

Combined

	F1	F1 (omitted)	abs.	rel. %
ComplementNB	79.53	67.60	-11.93	-15.01
LinearSVC	83.66	47.35	-36.31	-43.40
LSTMClassifier	84.58	47.03	-37.55	-44.40

Stability - Remarks/Improvements

- Generative classifier more stable than our discriminative ones
- **Problem:** LIME doesn’t scale well to **complex models**
- **Future:** Compare same architecture with different hyperparameters
  - **Example:** How many LSTM layers for a more stable prediction?
  - Use bootstrap significance tests
- Similarities to **Dropout layer** for neural models

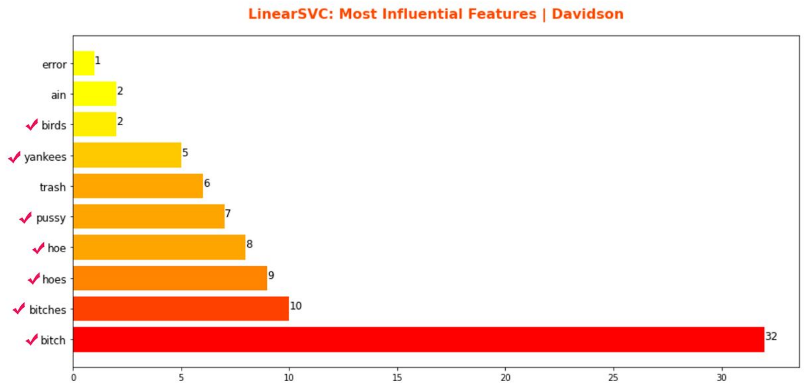
Most Influential Features

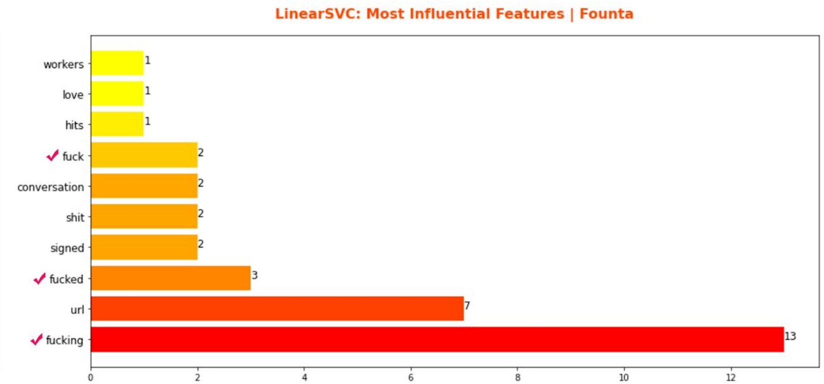
- Analysis made over **100 instances**
- Over each dataset, for the following classifiers:
  - Linear SVC
  - Complement NB
  - LSTM
- Two different statistics gathered:
  - MIF for **each decision** classifiers made
  - MIF for **wrong decisions** classifiers made

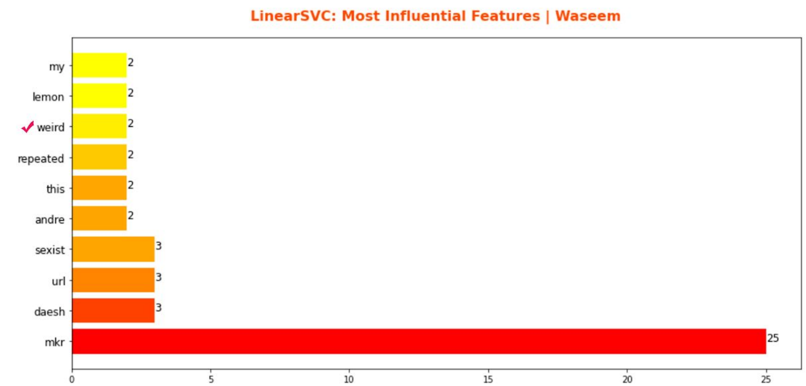
Most Influential Features

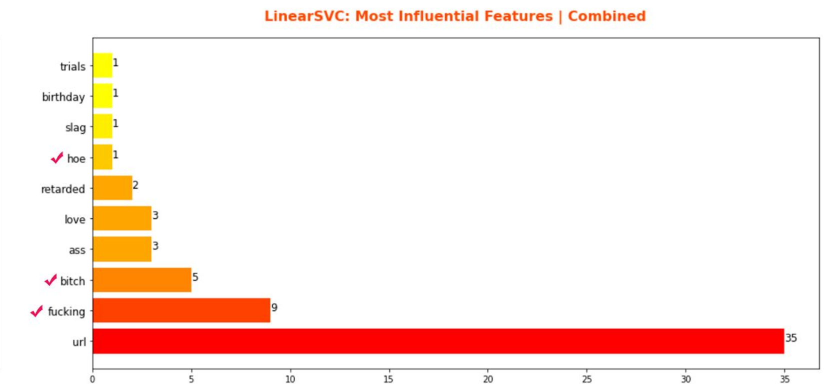
*Over **all** decisions*

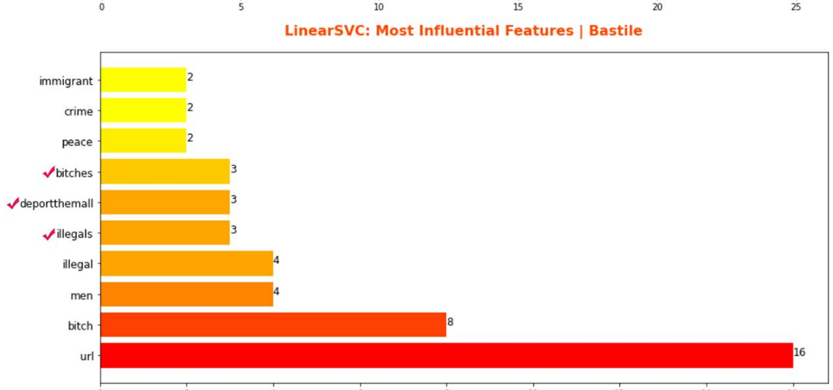
Linear SVC

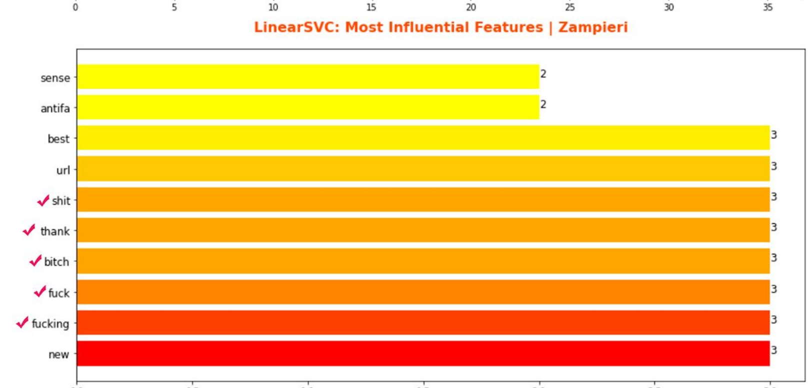








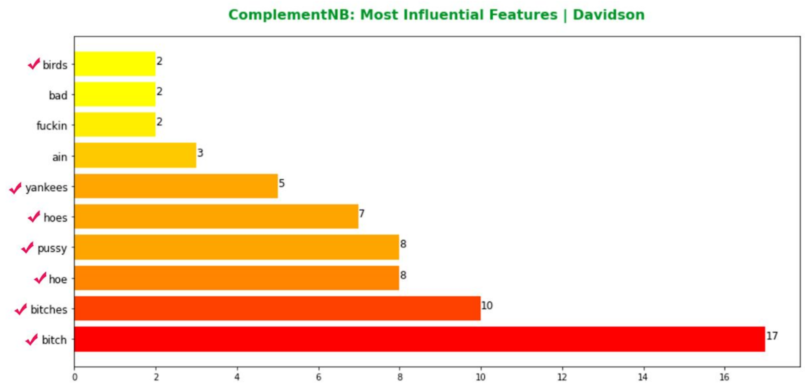


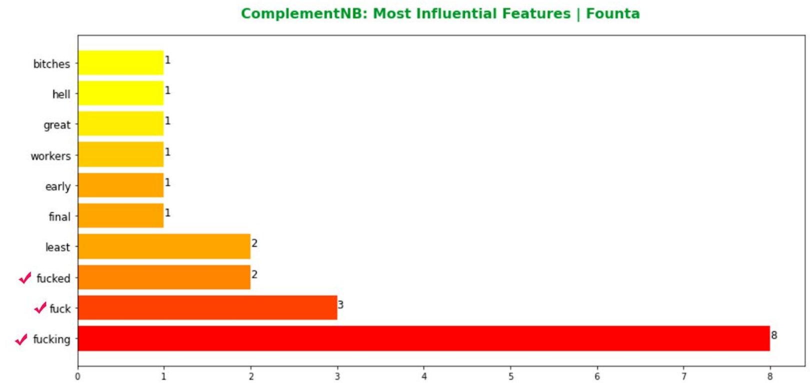


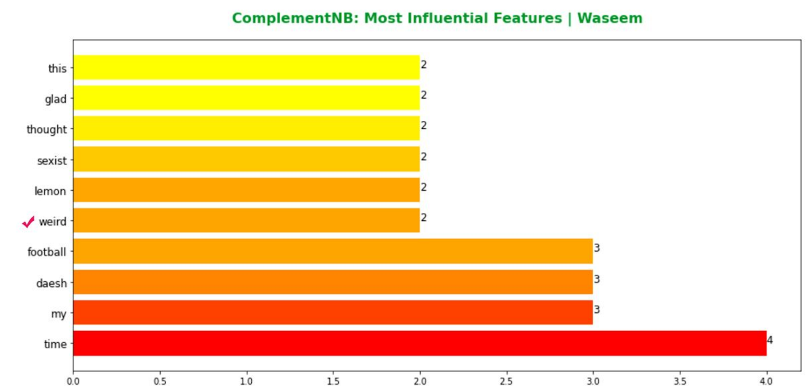
Most Influential Features

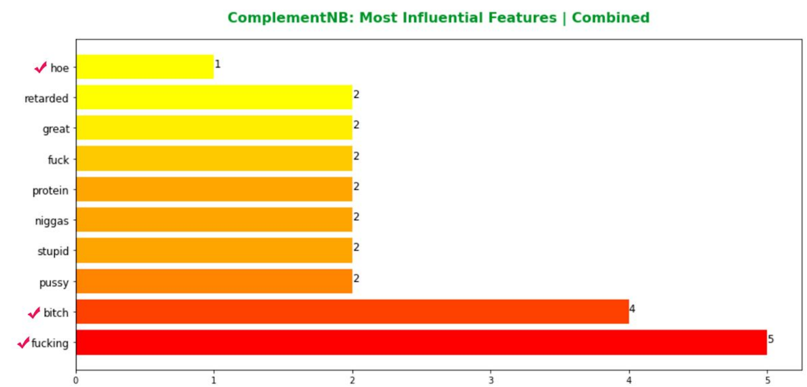
*Over **all** decisions*

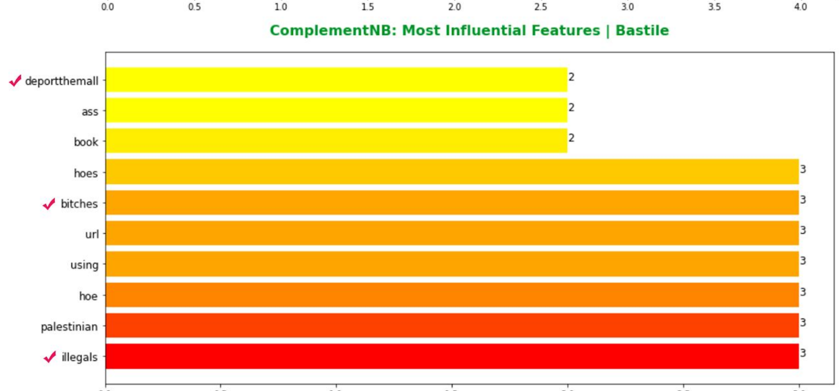
Complement NB

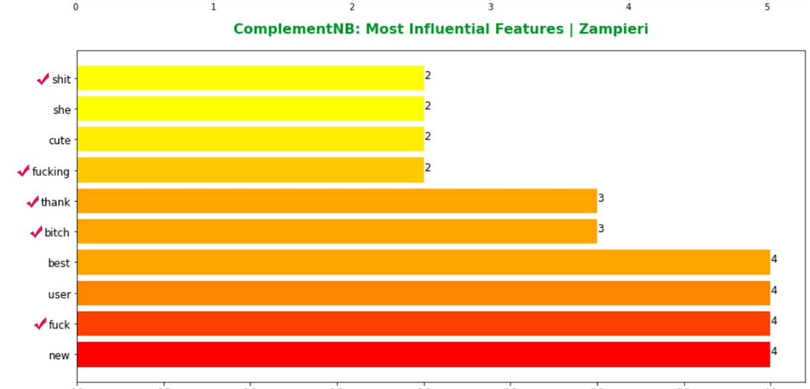








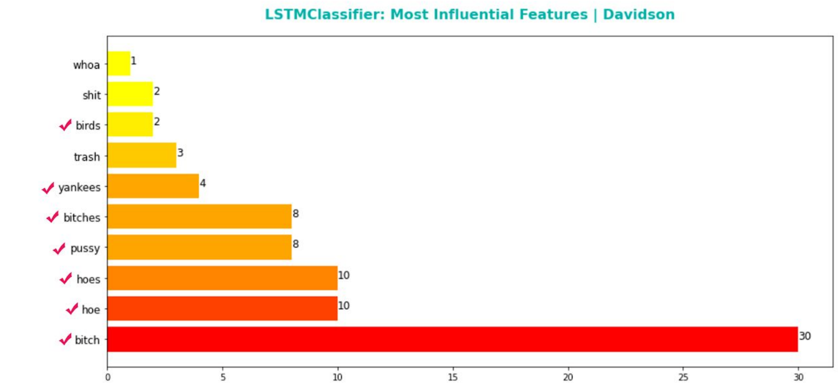


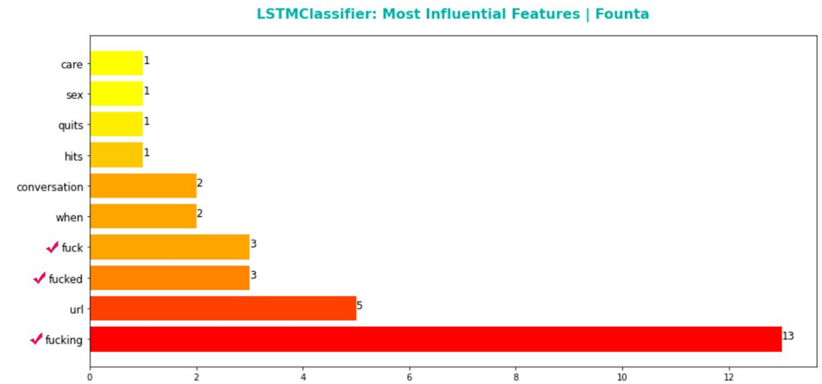


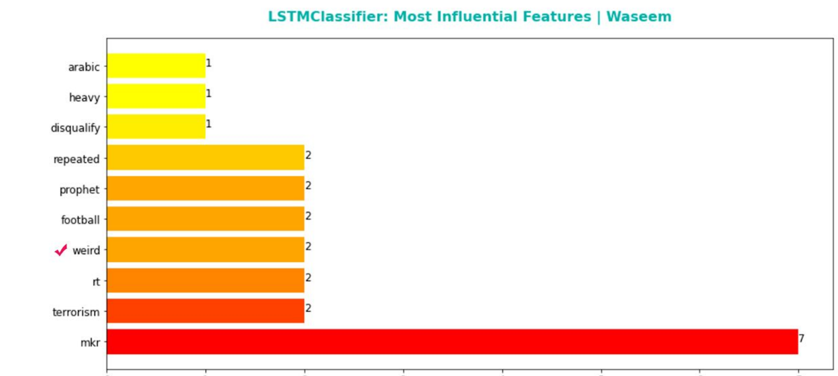
Most Influential Features

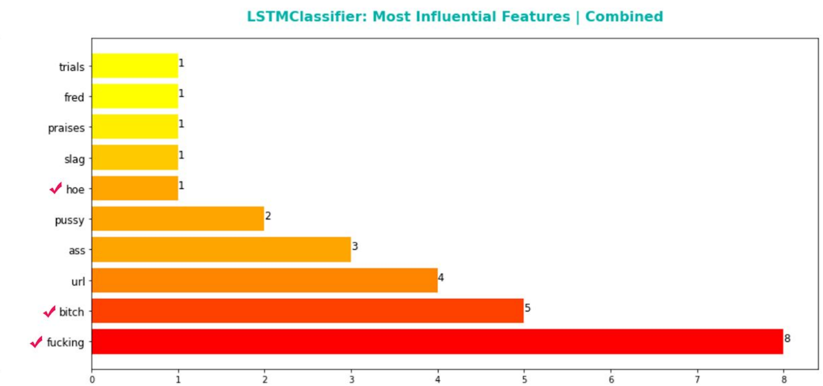
*Over **all** decisions*

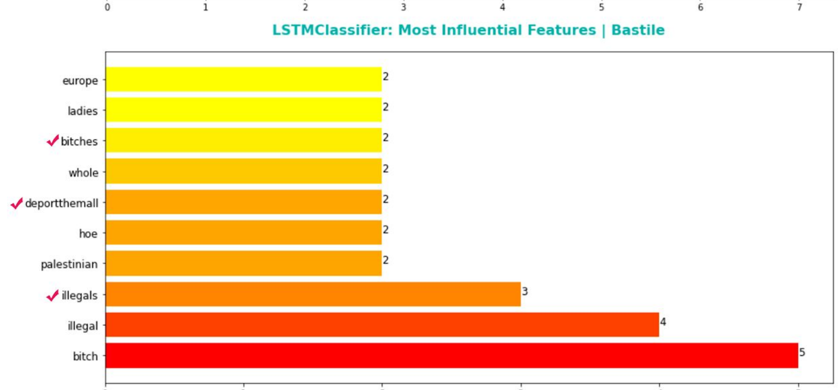
LSTM

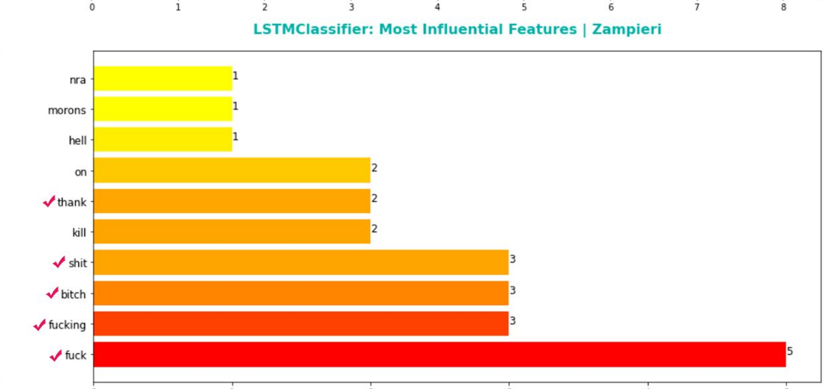








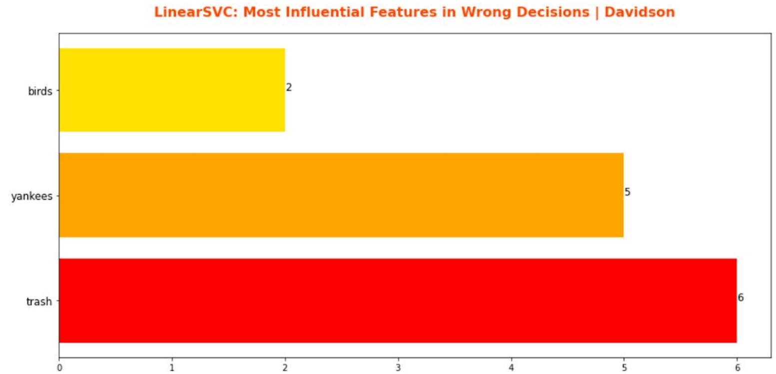


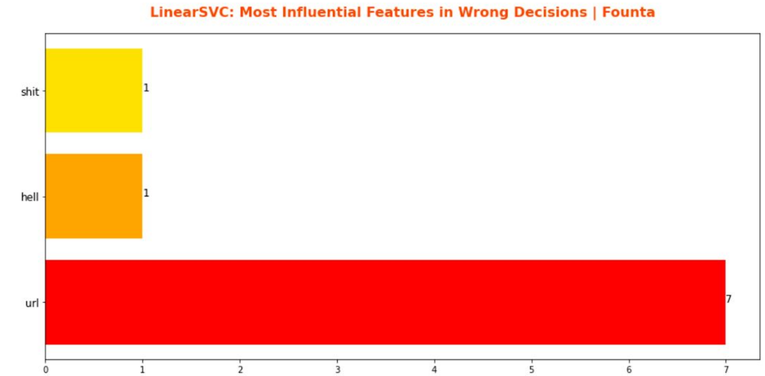


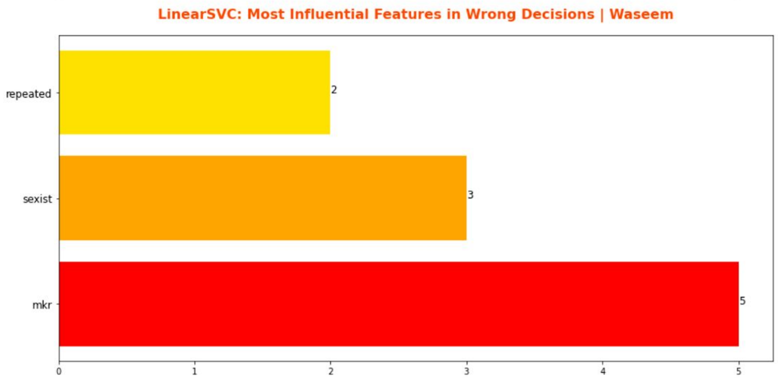
Most Influential Features

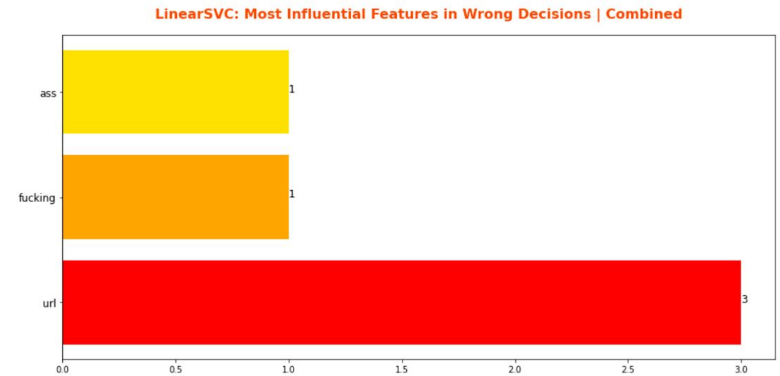
*Over **wrong** decisions*

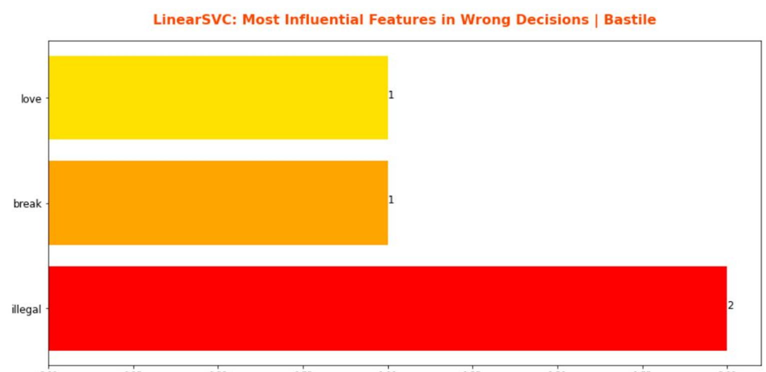
Linear SVC

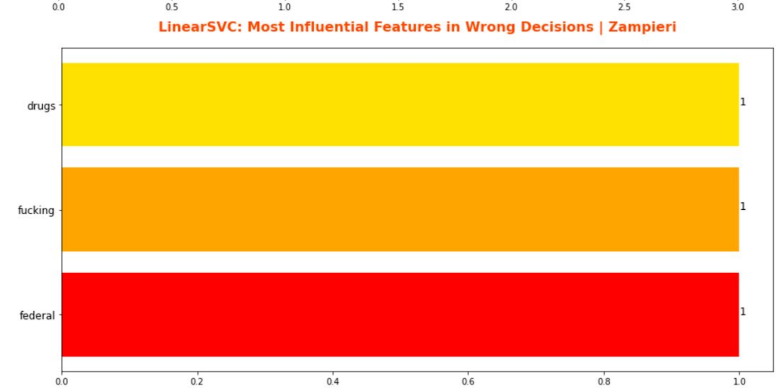








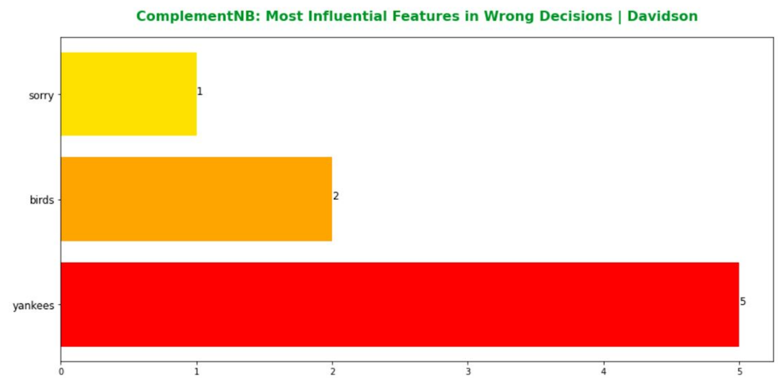


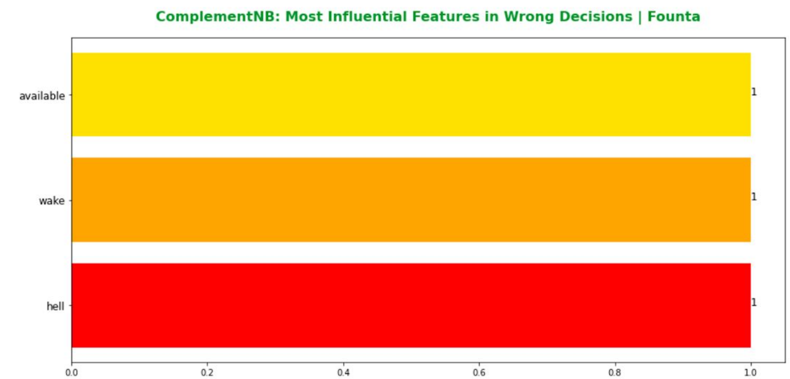


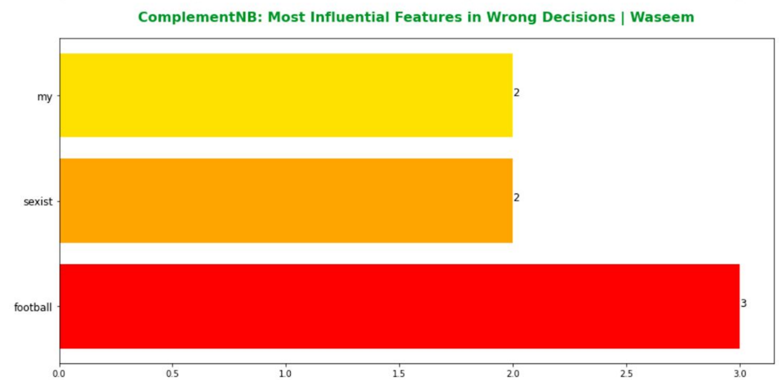
Most Influential Features

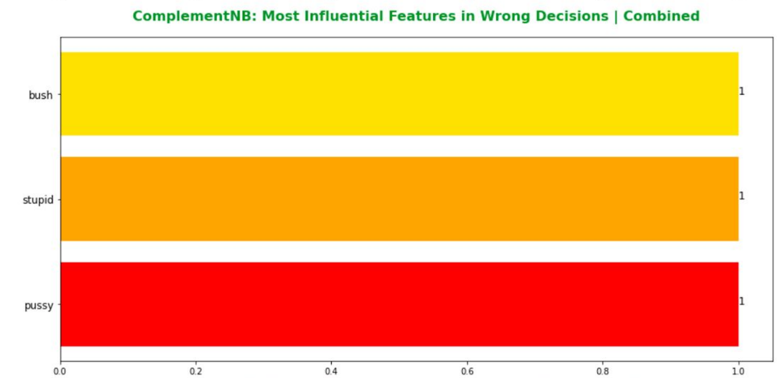
*Over **wrong** decisions*

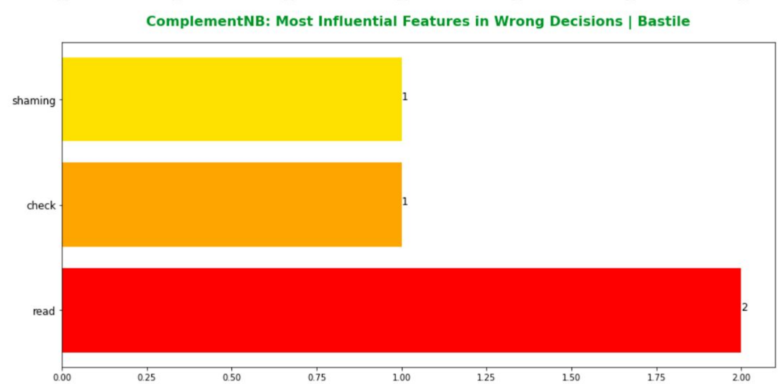
Complement NB

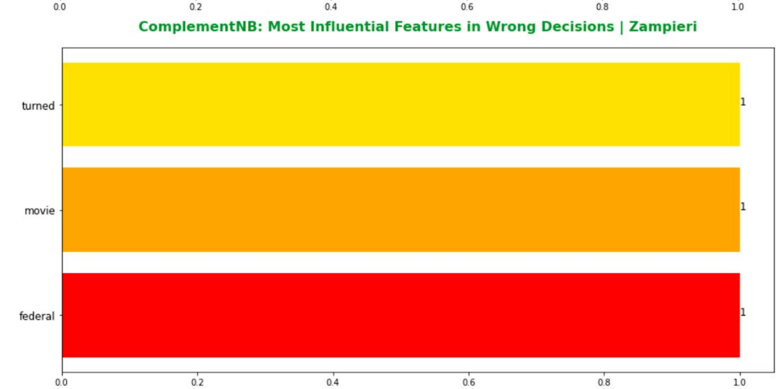








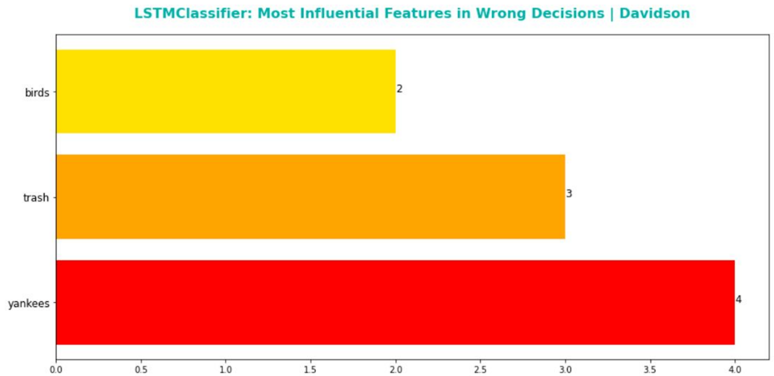


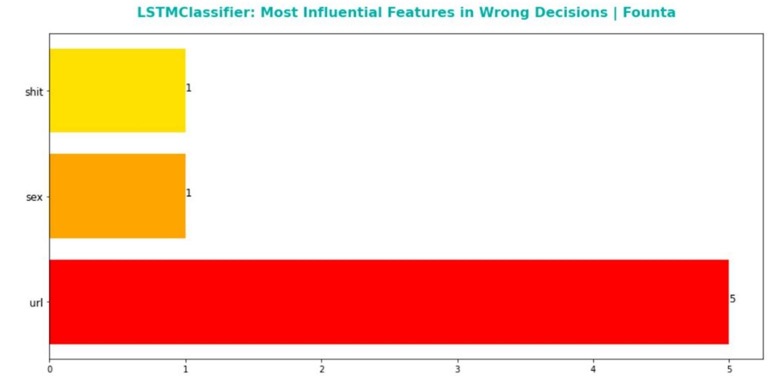


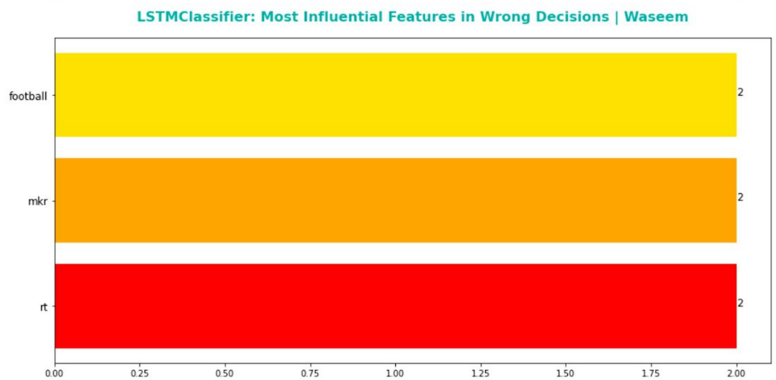
Most Influential Features

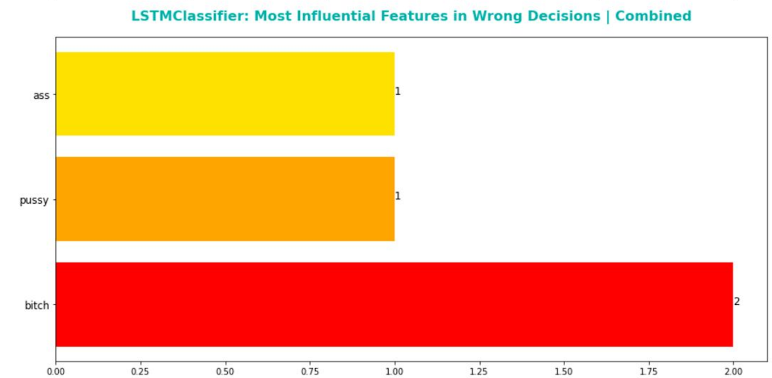
*Over **wrong** decisions*

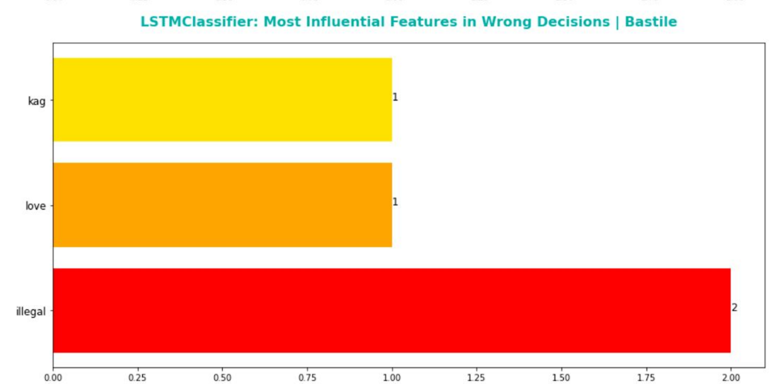
LSTM

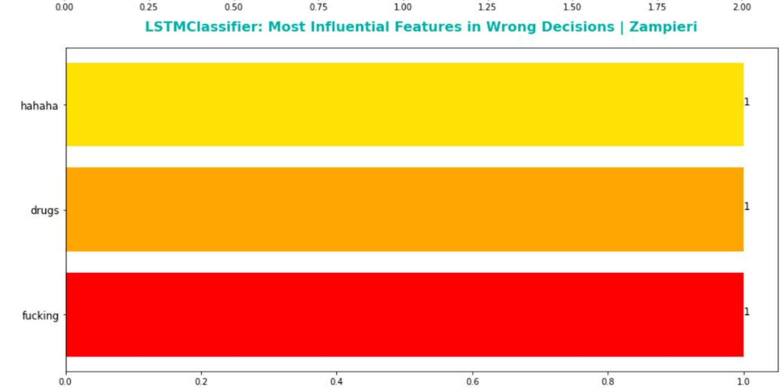








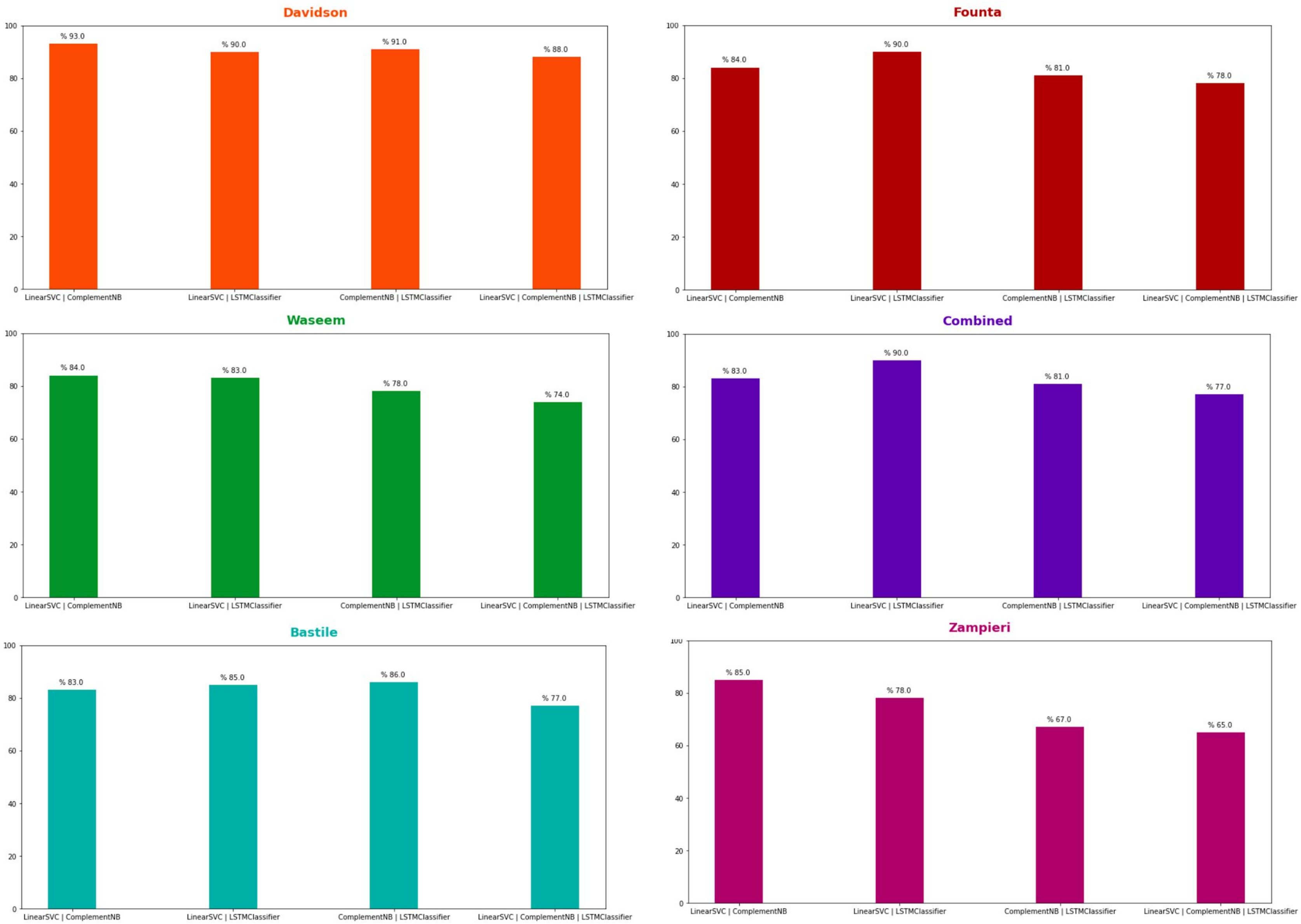




Comparison of Decisions

- Percentages for the combination of classifiers where **they made the same decision**
- Analysis made over **100 instances**
- Over each dataset, for all combinations of the following classifiers:
  - Linear SVC
  - Complement NB
  - LSTM

Comparison of Decisions



Problems & Conclusion

- Finding differences instead of similarities is hard
- Findings highly depend on the dataset
- Too many datasets and classifiers to choose from
  - Focused approach might be better
- Blackbox approach might be not as insightful

Questions

?