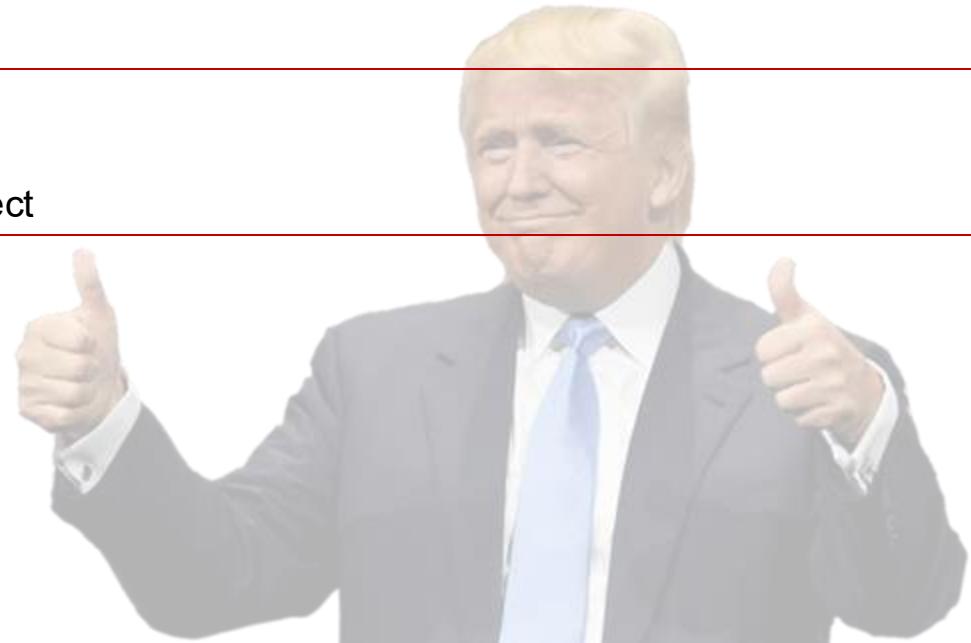




UNIWERSYTET WARSZAWSKI
Wydział Nauk Ekonomicznych

Donald Trump related tweets during 2022 U.S. election

Sentiment Analysis Final Project



Warsaw, 21st of January 2026

Cezary Kuźmowicz, Michał Sucharzewski

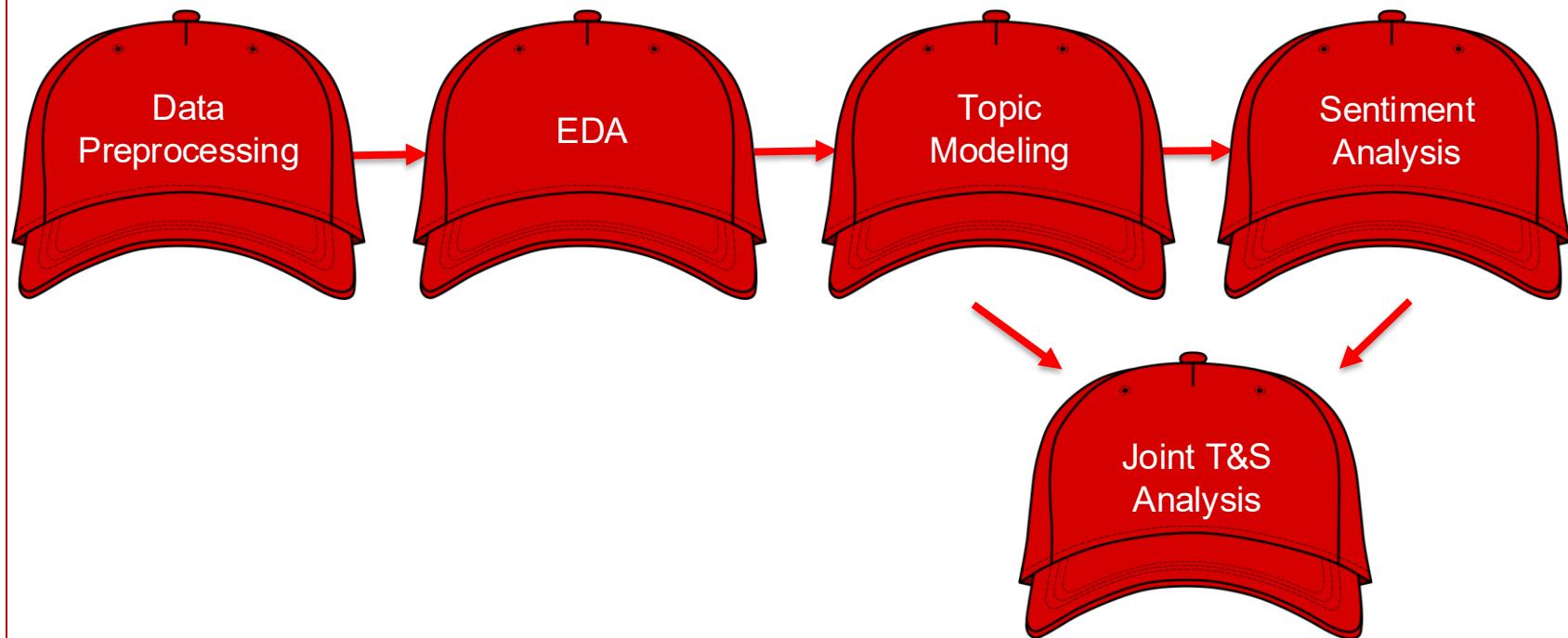


Data

- **Source:** Kaggle (Tweets related to hashtags #DonaldTrump and #Trump).
- **Scale:** Initially around 1mln, after language detection over **500k English tweets**.
- **Timeframe:** October-November 2020 (Leading up to election).
- **Features Used:** Tweet text, creation date
- **Filtering:** Focused on English language tweets; removed retweets to analyze original opinions only.

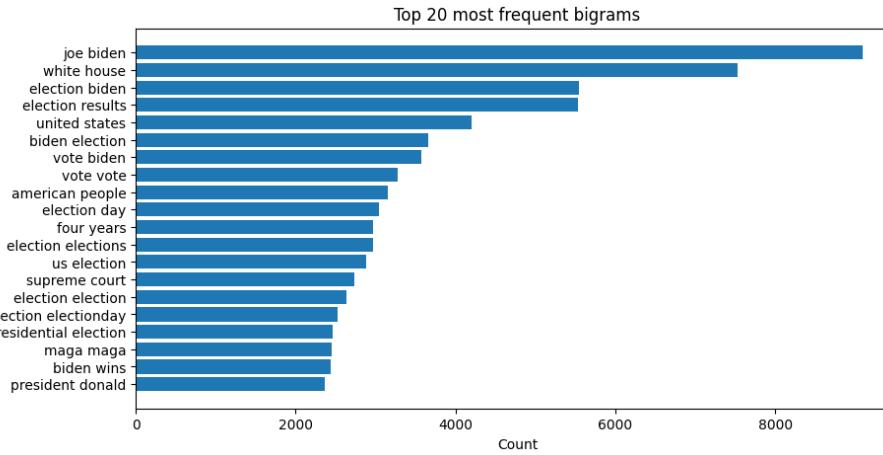
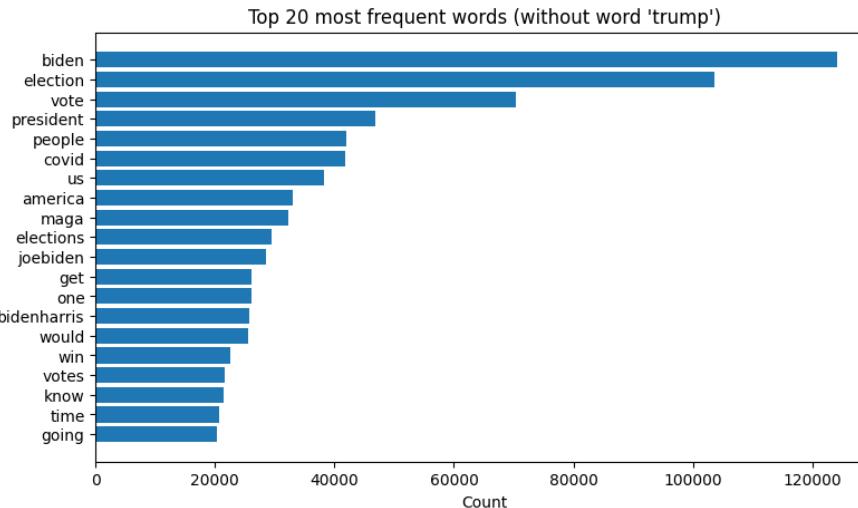


Workflow





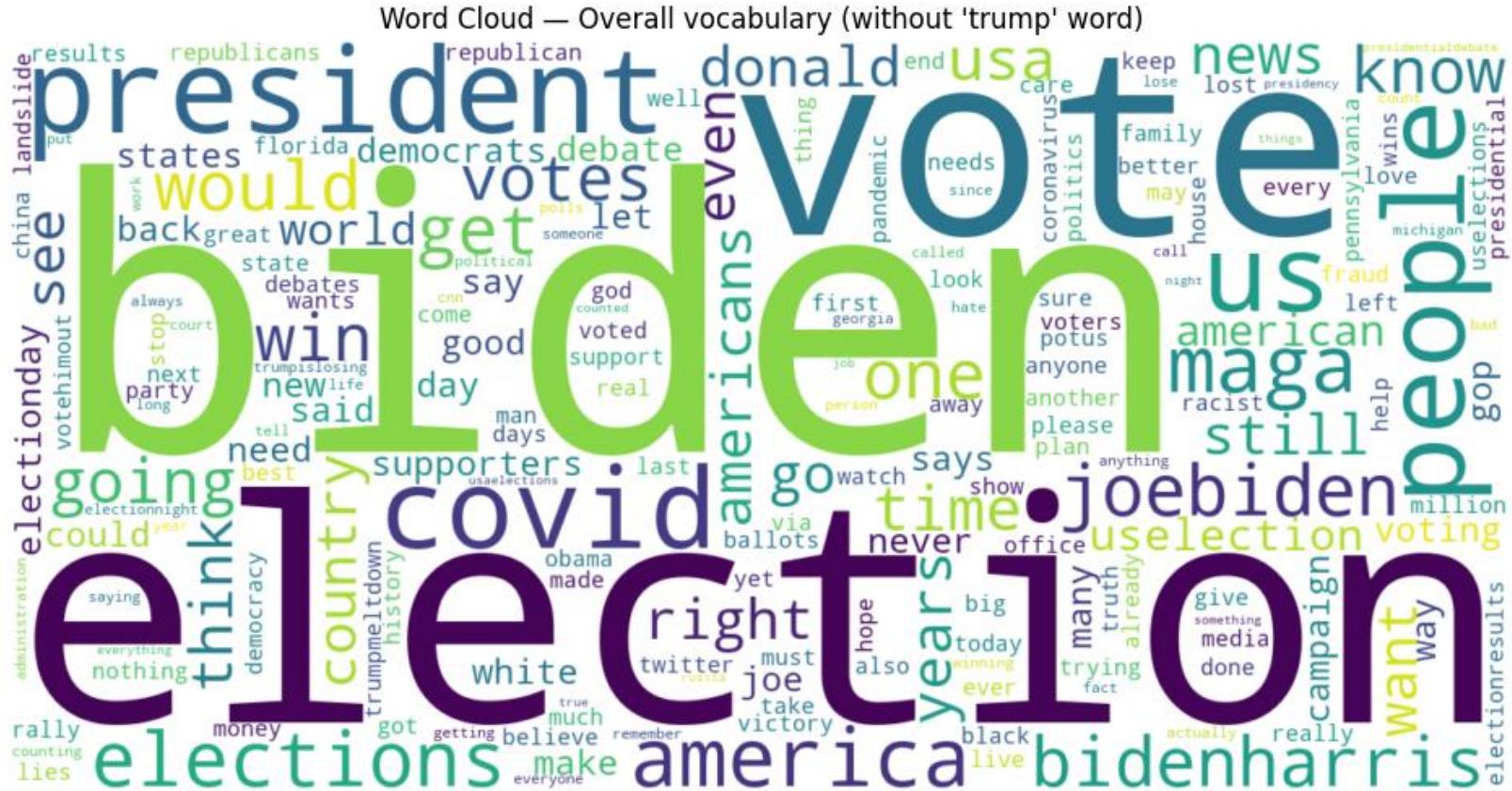
Data Preprocessing



- Basic cleaning and normalization
- Stop Words Removal
- Dual-Cleaning Strategy for Topic Modeling and Sentiment Analysis

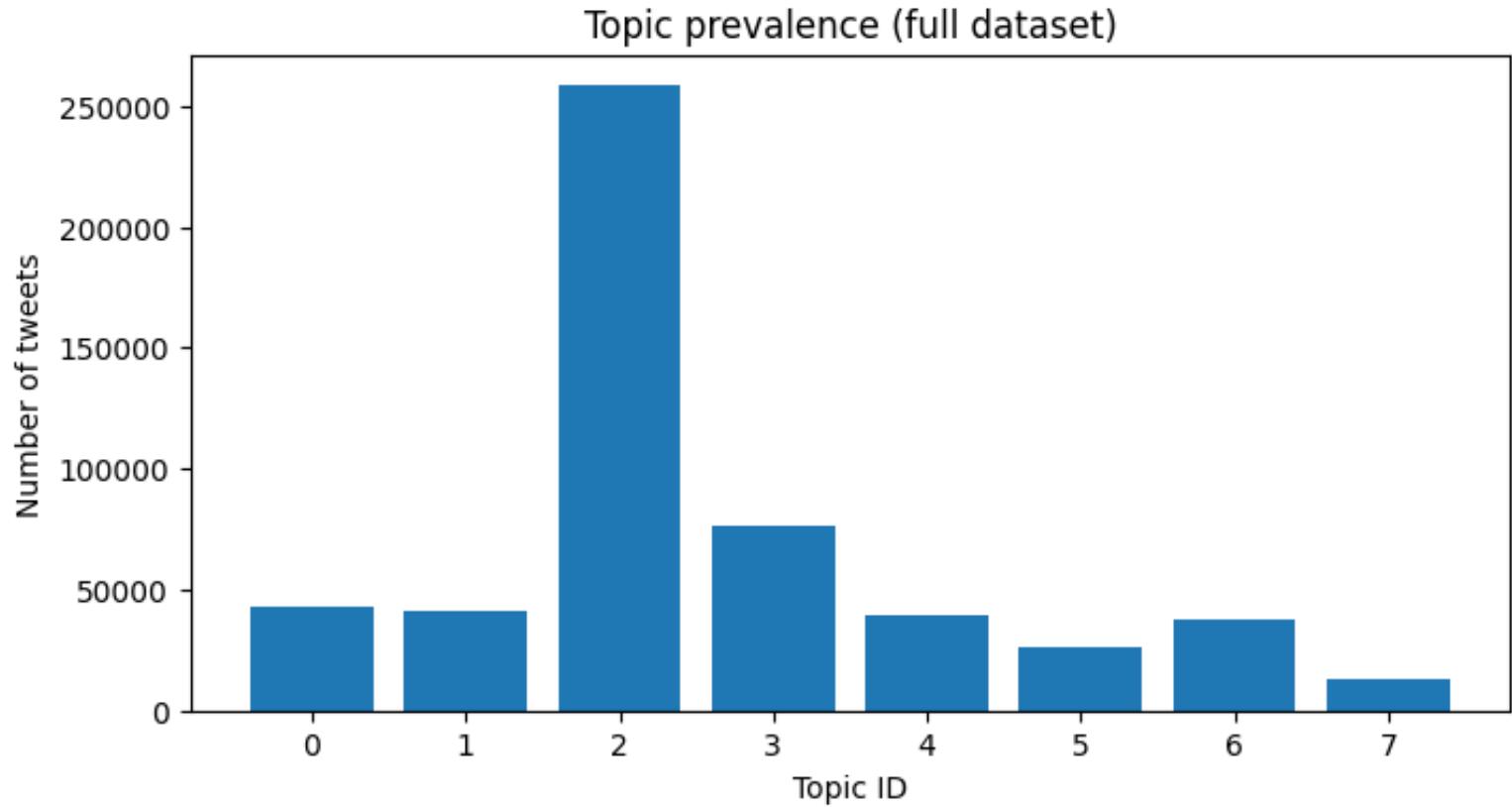
	n_words	n_chars
count	534344.000000	534344.000000
mean	25.489001	151.385476
std	12.938555	72.881908
min	3.000000	10.000000
25%	15.000000	89.000000
50%	24.000000	146.000000
75%	36.000000	219.000000
max	68.000000	291.000000

Keywords Map



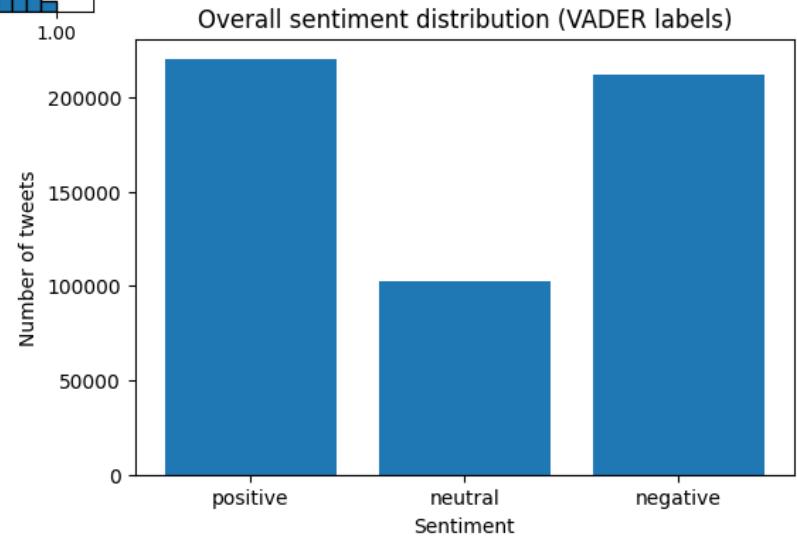
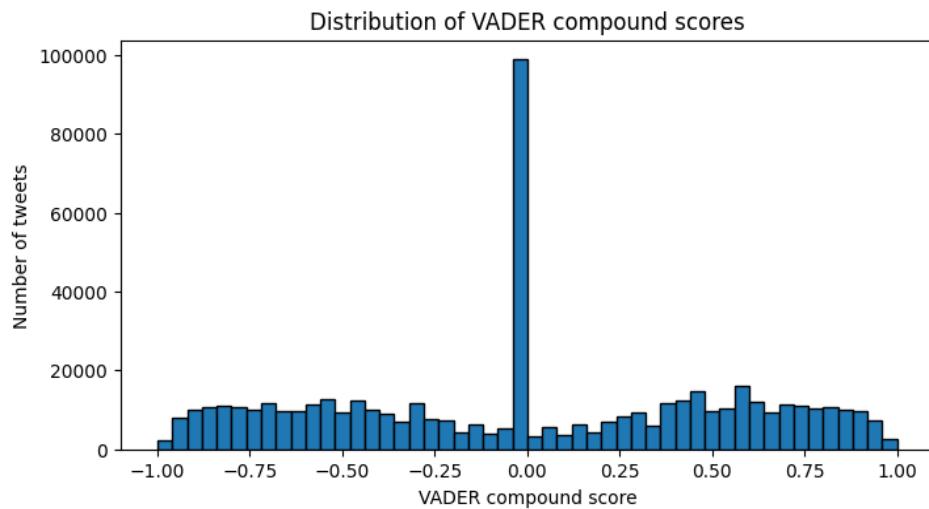


Topic Modeling – TF-IDF & NMF





Sentiment Analysis – Vader

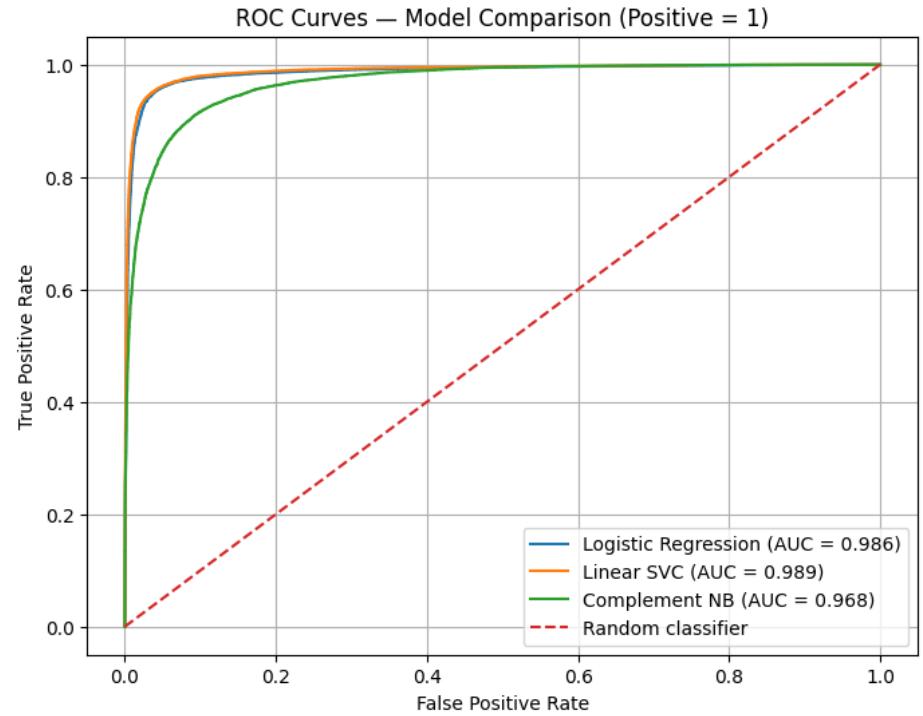




Supervised Sentiment Analysis

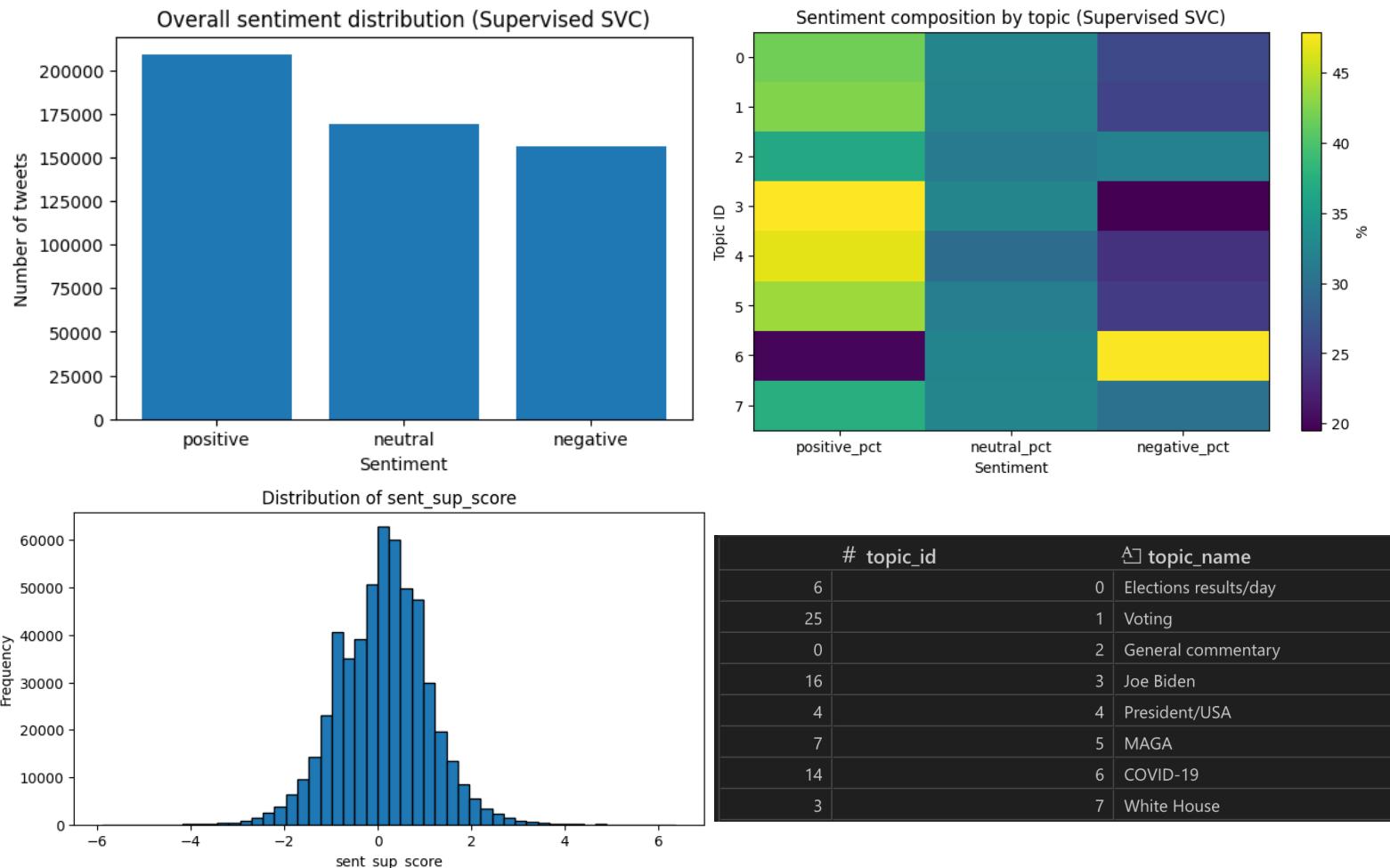
	model	best_cv_f1	test_f1	test_acc	test_roc_auc
1	TF-IDF + LinearSVC (tuned)	0.9568	0.9571	0.9571	0.9887
0	TF-IDF + LogReg (tuned)	0.9549	0.9557	0.9556	0.9863
2	TF-IDF + ComplementNB (tuned)	0.9080	0.9072	0.9077	0.9684

- Logistic Regression and LinearSVC performed almost the same – **the second one was slightly better**
- Complement Naïve Bayes perfromed the worst of all models
- On that step we created only "**positive/negative**" options – after applying to whole dataset we created "**neutral**" label



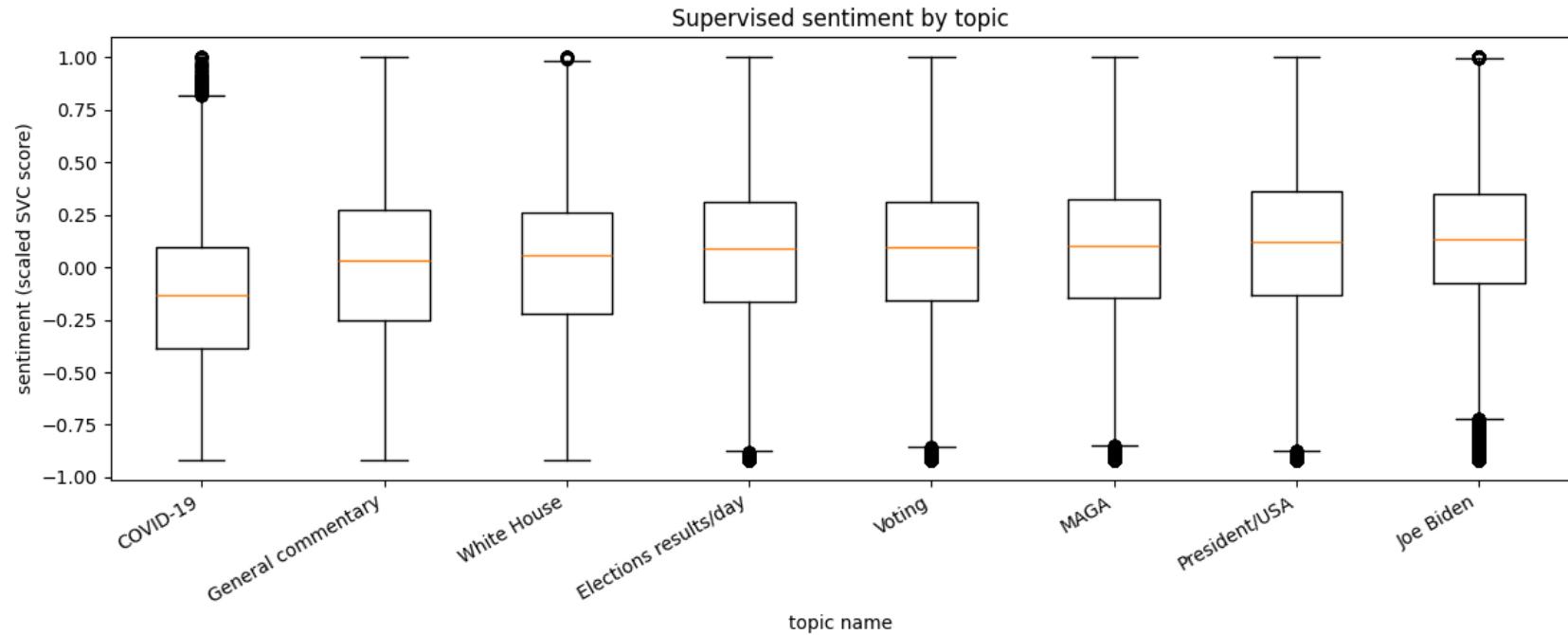


Sentiment x Topic Joint Analysis





Sentiment x Topic Joint Analysis





Key takeaways

- The most positive topic was... "**Joe Biden**" - likely due to supportive campaigning and 'hopeful' language used by supporters, or the sarcastic usage of hashtags by opponents
- The clear and obvious winner in negative sentiment was **COVID-19**. From well known reasons it got definitely the lowest sentiment among all topics
- Dataset was **really large** and all topics were **widely distributed** regarding the sentiment





UNIWERSYTET WARSZAWSKI
Wydział Nauk Ekonomicznych

Thank you for attention!



Warsaw, 21st of January 2026

Cezary Kuźmowicz, Michał Sucharzewski