Embeddings-Based Clustering for Target Specific Stances

Ammar Raşid

github.com/AmmarRashed/UnsupervisedStanceDetection

Prof. Dr. Tanju Erdem
Dr. Reyyan Yeniterzi

ÖZYEĞİN UNIVERSITY

Outline

Introduction

What is a stance?

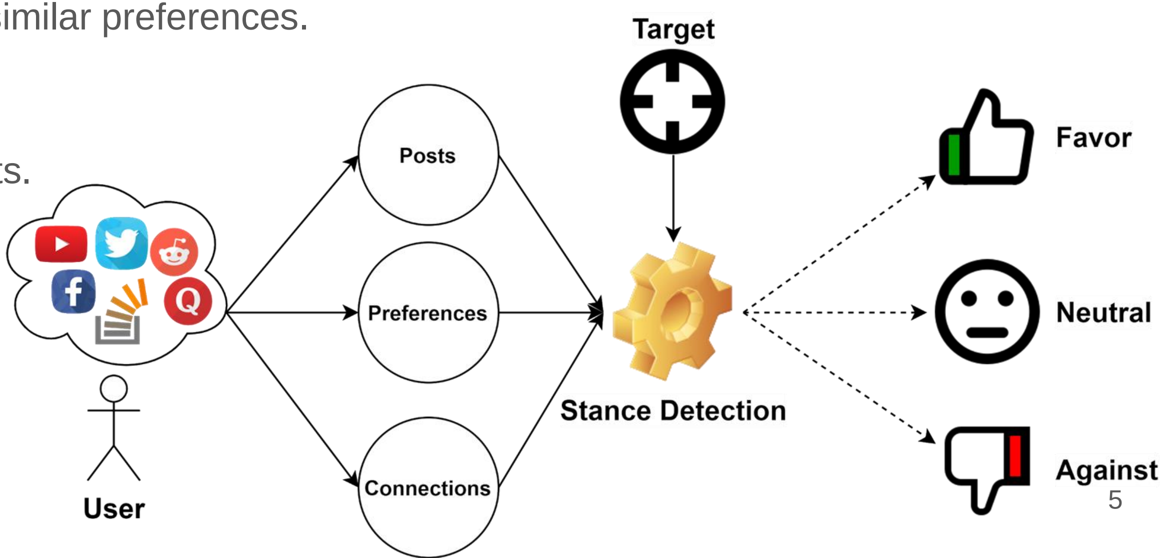
- **Definition:**
  - The position of a user towards an issue or an entity.
- Single:
  - **PRO:**
    - Take the Red Pill, sleep well, WE WILL *MAKE AMERICA GREAT AGAIN!*
  - **ANTI:**
    - "*Make America Great Again*" is actually a dog whistle to fascism and white supremacy.
- Multiple Targets:
  - I am a **Liverpool** fan but today my heart goes out to **Newcastle** fans
- **Stance is not sentiment**

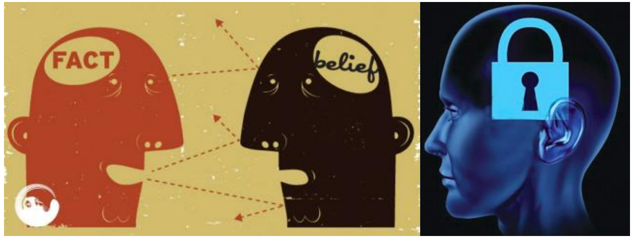
+ stance - sentiment: *I feel bad that he lost*

- stance + sentiment: *I am so glad he died*

Why stance detection?

- **Views collection:**
  - Useful for marketing and polling.
- **Recommendation:**
  - Similar stances could imply similar preferences.
- **Population Analysis:**
  - Dissecting polarization targets.



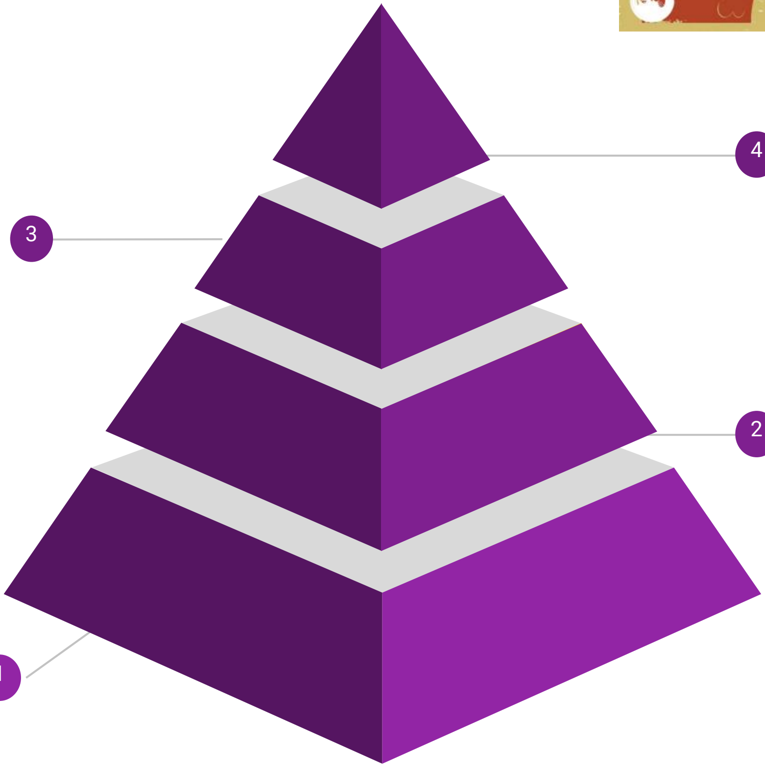


4

Biased Assimilation
welcome confirming evidence less critically than disconfirming ones.

3

Echo chambers
closed communities share the same information sources.

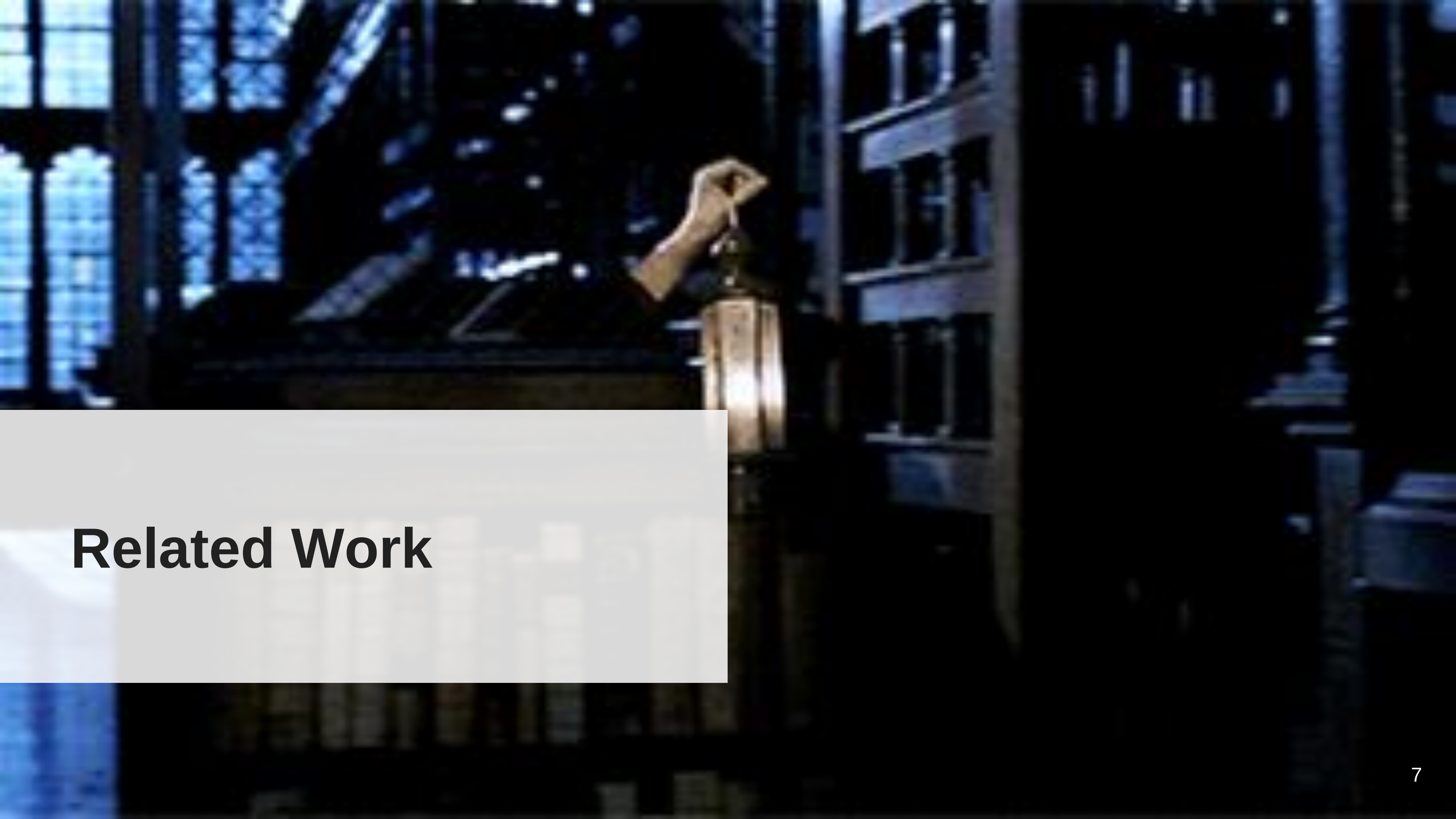


2

Social Influence
once people are in groups, they influence each other.

1

Homophily
Birds of a feather, flock together



Related Work

Supervised Stance Classification

- **Pre-trained word embeddings:**
    - **MITRE** (Zarrella & Marsh 2016)**:** 256-dimensional SkipGram -> LSTM -> RLU
- **Bidirectional Conditional Encoders (BCE):** Augenstein et. al 2016
    - **Cross-Target** (Xu et. al 2018)**:** CrossNet: BCE + self-attention mechanism.
    - **DMAN** (Wei et. al 2018)**:** dynamic memory augmented network, BCE + attention + memory modules.
        - Memory networks can provide snippets of the stance evidence in documents
- **Sentiment Information:** “Stance detection via sentiment information and neural network model, Sun et. al (2018)”
    - jointly learn sentiment and stance information.
    - Sentiment information is fed to an auxiliary shared LSTM between stance and sentiment data.

Stance Detection in Twitter

- **Feature stratification on Twitter:**
  - **Text:** e.g. n-grams.
  - **Interaction:** e.g. retweets, replies, mentions ...etc
  - **Preference:** e.g. likes
  - **Connections:** e.g. followers, friends ...etc
- **Model:**
  - Most competitive models used **SVM** with retweeted accounts as features.
- **Your stance is exposed!**: Aldayel & Magdy 2019
  - Silent users can still have their stance detected using only their connections (i.e. F1 = 71.39)

Unsupervised Stance Detection

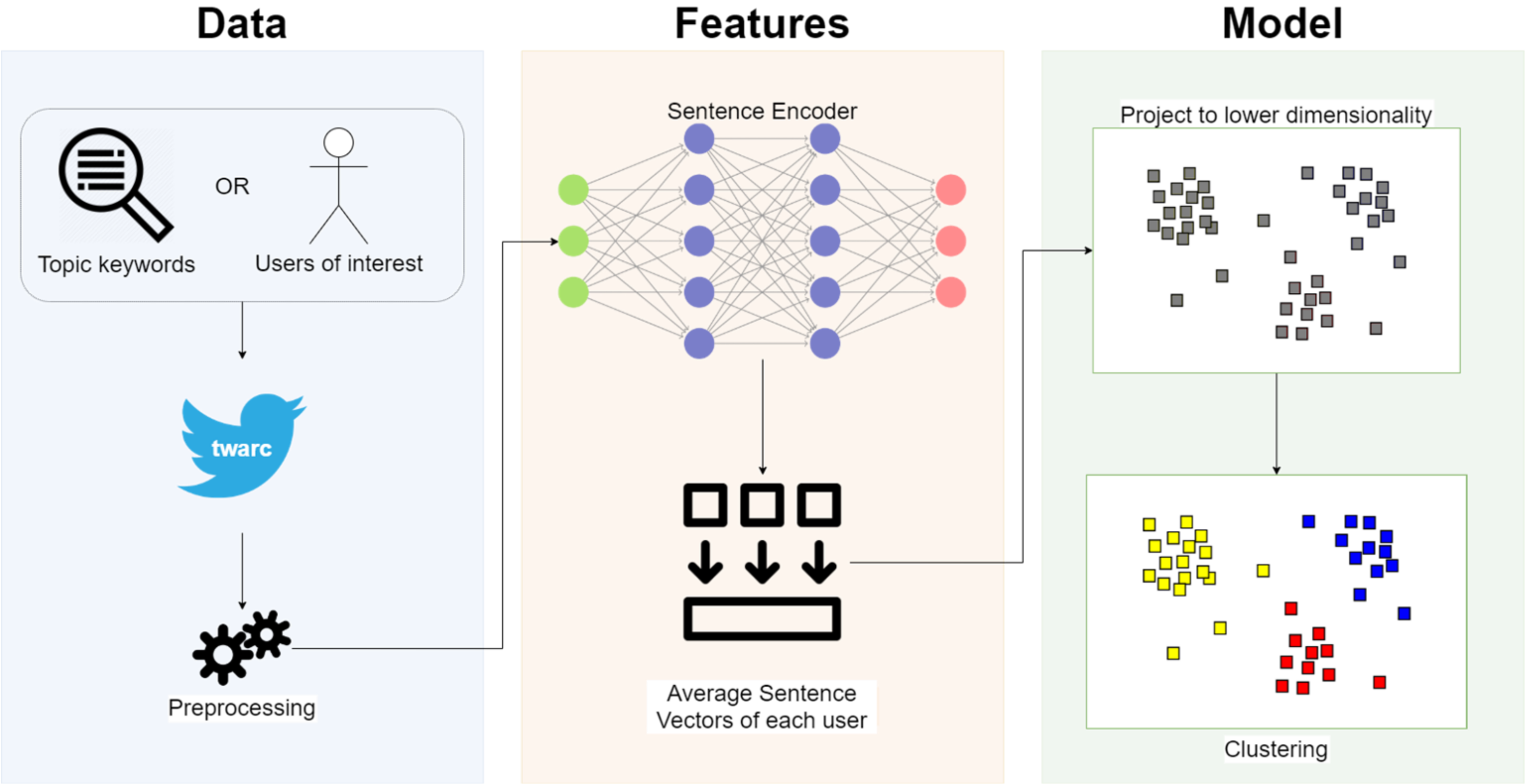
- **Text-based:** “*Detection of Stance-Related Characteristics in Social Media Text*, Simaki et. al 2018”
  - 21 stance-related text characteristics, e.g. contrariety, hypotheticality, and necessity constructions.
  - Project using PCA.
  - Organize projected vectors into 6 clusters of different writing styles.
- **Interaction-based:** “*Unsupervised User Stance Detection on Twitter,* Darwish et. al 2019”
  - Hashtags, and retweeted accounts as features.
  - Project using UMAP.
  - Cluster with Mean-shift.

Limitations of previous best Unsupervised Stance Detection

- Requires platform-specific features.
- Requires large amount of data for training.
- Does not consider the hierarchy of stances.
- Requires specification of the number of clusters.

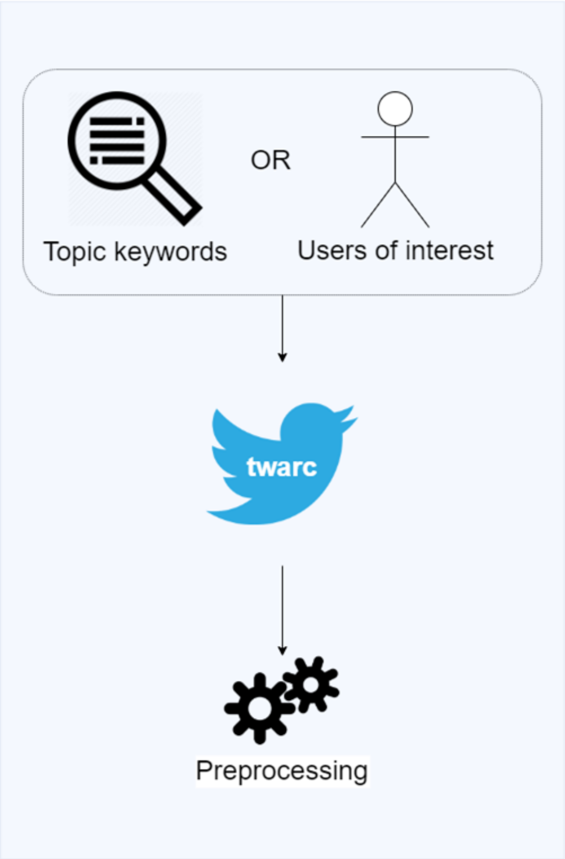
Methodology

Framework


Data
Features
Model
OR
Topic keywords
Users of interest
twarc
Preprocessing
Sentence Encoder
Average Sentence
Vectors of each user
Project to lower dimensionality
Clustering

Data

- Data Collection
  - Keywords
  - Usernames
- Cleaning
  - Converting camel-cased hashtags to space-separated phrase.
  - Removal of all links, and user mentions.
  - Removal of all non-letter characters and punctuation.
  - Replacement of all numbers to the word “number”.
  - Case folding, where we lowercase all letters.
- Filtering
  - Min 3 tokens per tweet
  - Min 3 unique tweets per user
  - Accounts suspended within 8 months are assumed Bots



Datasets: UEFA

- **Topic:** UEFA Super Cup, Istanbul 14th of August 2019
- **Date span:** Aug. 8-17, 2019
- **Language:** English
- **Labels:** Liverpool fan (*LFC*) vs Chelsea fan (*CFC*)
- **Nature:** More text, less #Hashtags and Retweets

Annotations:
match keywords
in user bios

Target	Keywords	Users	Tweets	Unique Tweets
CFC	KTBFFH, Chelsea FC, #CFC, Chelsea fan	3,821	46,369	17,672
LFC	YNWA, Liverpool FC, #LFC, Liverpool fan	8,141	99,582	29,192
	Total	11,962	145,978	46,213
	Other	24,694	166,678	166,678

Datasets: Trump

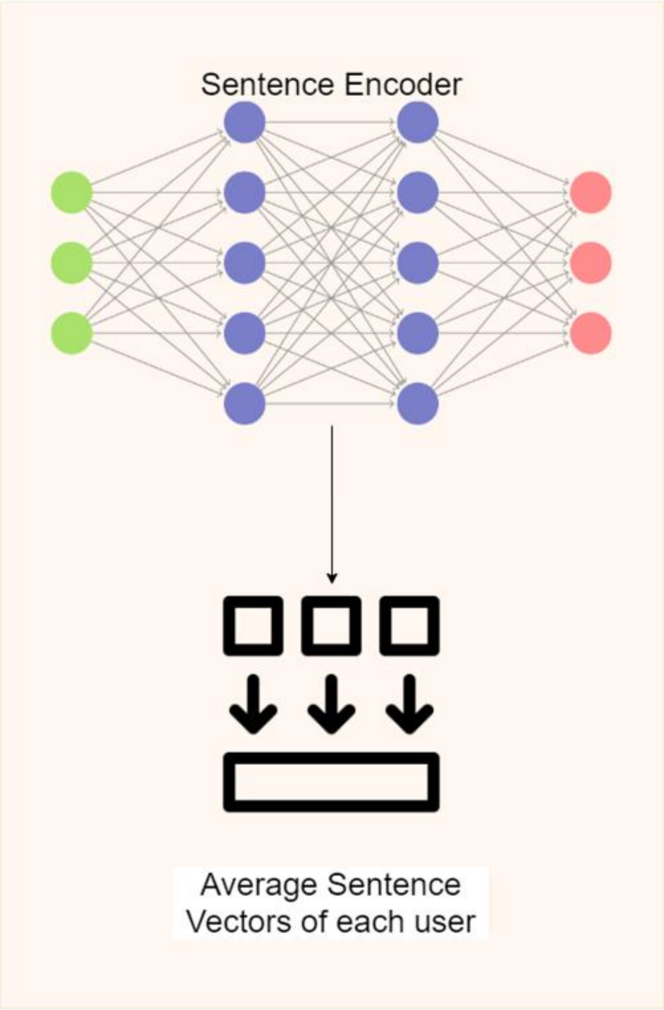
Unsupervised User Stance Detection on Twitter, Darwish et. al 2019

- **Topic:** USA midterm presidential elections
- **Date span:** Oct. 25-27, 2018
- **Language:** English
- **Labels:** Pro Trump (*PRO*) vs Anti Trump (*ANTI*)
- **Nature:** Too much #Hashtags & Retweets!

Target	Keywords	Users	Tweets	Unique Tweets
PRO	#MAGA	5,803	339,886	22,819
ANTI	#resist, #resistance, #impeachTrump, #theResistance, #neverTrump	4,788	293,449	25,749
	Total	10,591	311,335	43,836

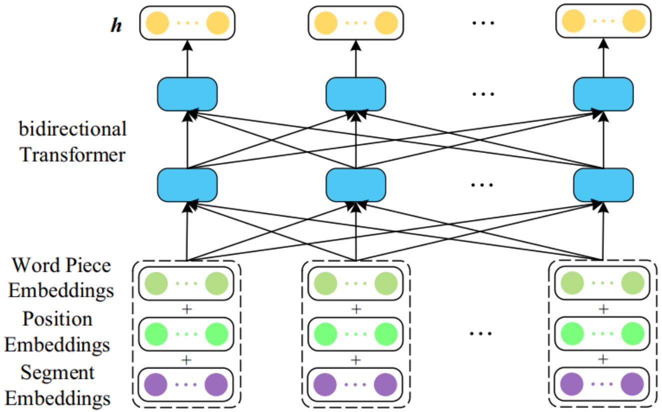
Features

- **Select:**
  - tweets containing stance targets.
- **Encode:**
  - tweets using a pre-trained sentence encoder.
- **Average:**
  - tweet vectors per user to obtain a single vector per user.

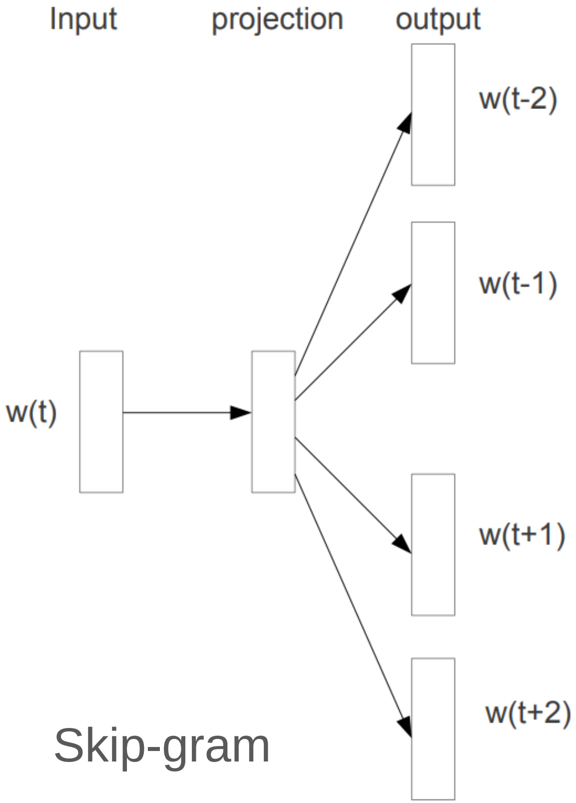


Word Embeddings

- Vector representation of words
  - Similar words have similar contexts.
  - Vectors of similar words are closer to each other.
- Word2Vec: e.g. CBOW, Skip-gram
- Transformer Models: e.g. BERT

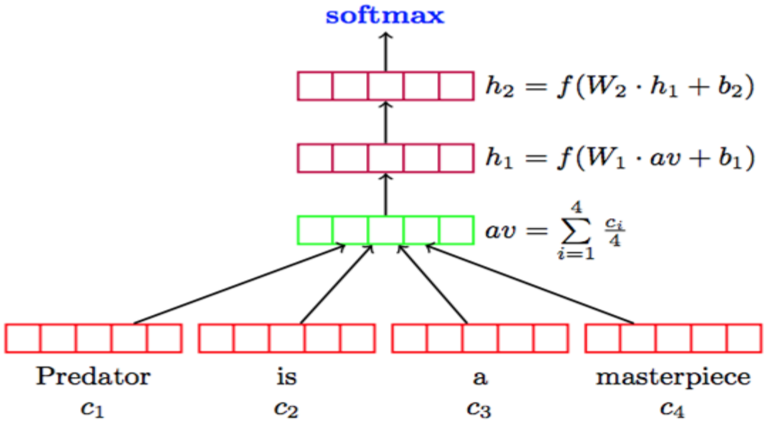


BERT

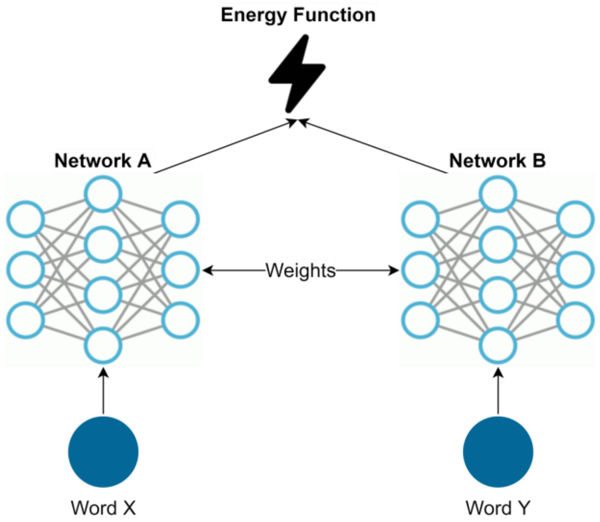


Distributed Representations of Words and Phrases and their Compositionality (Mikolov et.al. 2013)
From static to dynamic word representations: a survey (Wang et.al. 2020)

Sentence Encoders



- Universal Sentence Encoder (USE):
  - Average input embeddings.
  - Pass through feed-forward DNN.

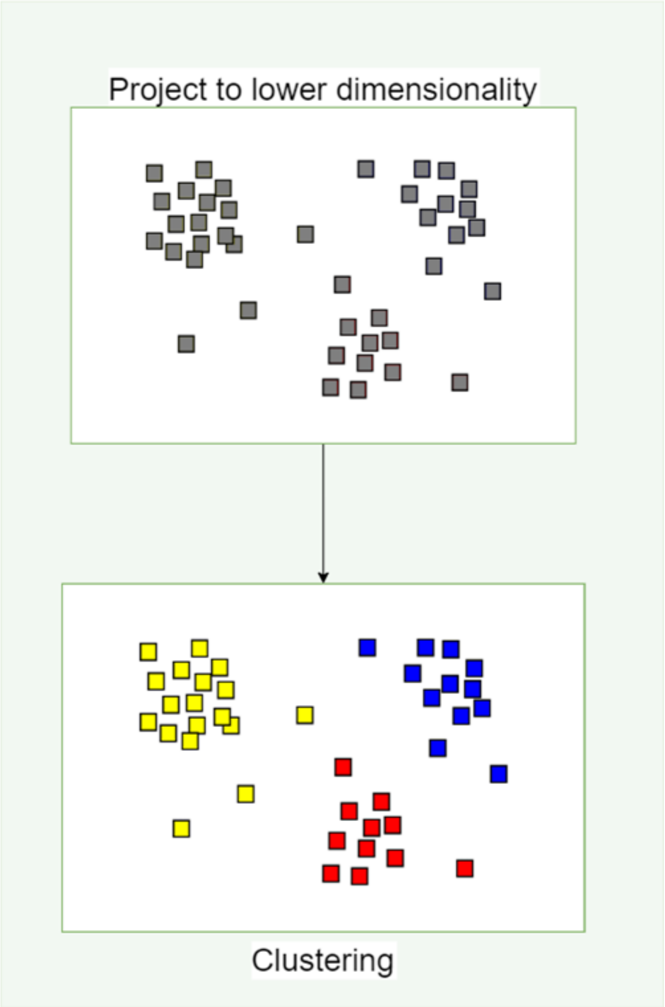


- Sentence-BERT:
  - Modify BERT to use Siamese Networks.
  - Siamese networks can be twins and triplets.

DAN: Deep Unordered Composition Rivals Syntactic Methods for Text Classification (Iyyer et.al. 2015)
USE: Cer et.al. 2018: https://tfhub.dev/google/universal-sentence-encoder/4
SBERT: Reimers & Gurevych https://github.com/UKPLab/sentence-transformers

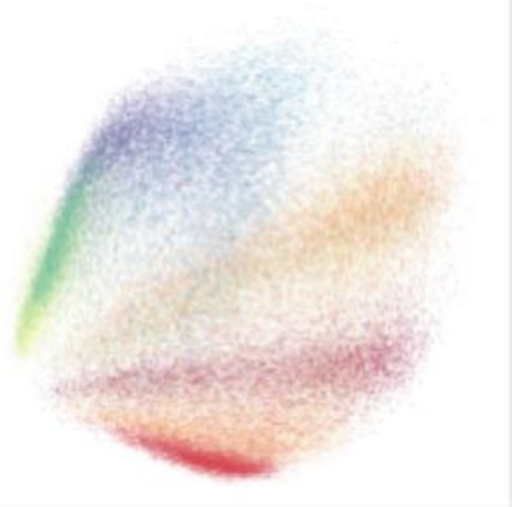
Model

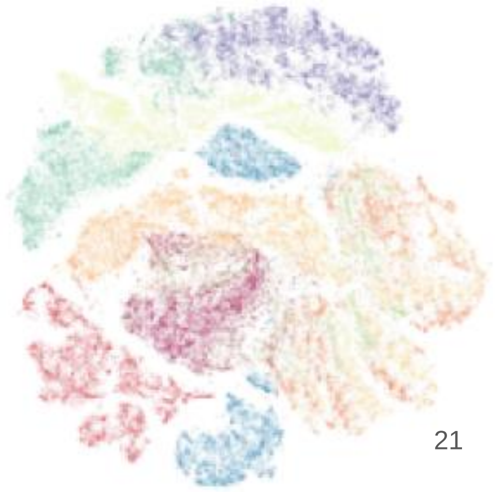
- **Project:**
    - user vectors onto a lower dimensional space.
- **Cluster:**
    - projected vectors into clusters of stances.



Vector Projection

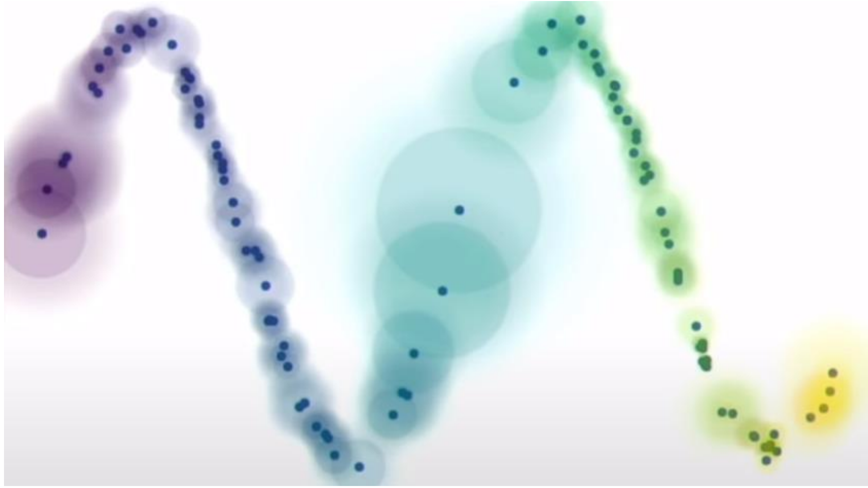
- **Taxonomy:**
    - **Matrix Factorization**: e.g. PCA, LDA
    - **Neighbor Graphs**: e.g. t-SNE, UMAP
- **Principal Component Analysis (PCA):**
    - Assumes latent correlation between variables.
    - Projects onto a space of an uncorrelated variables.
    - Maximize variance while preserving **long** pairwise distances.
    - Poor with non-linear manifold structures.
- **t-distributed stochastic neighbor embedding (t-SNE)**
    - Centers a gaussian distribution P about a data point X.
    - Measures the density of all other data points under that distribution.
    - Focuses on preserving **short** pairwise distances.

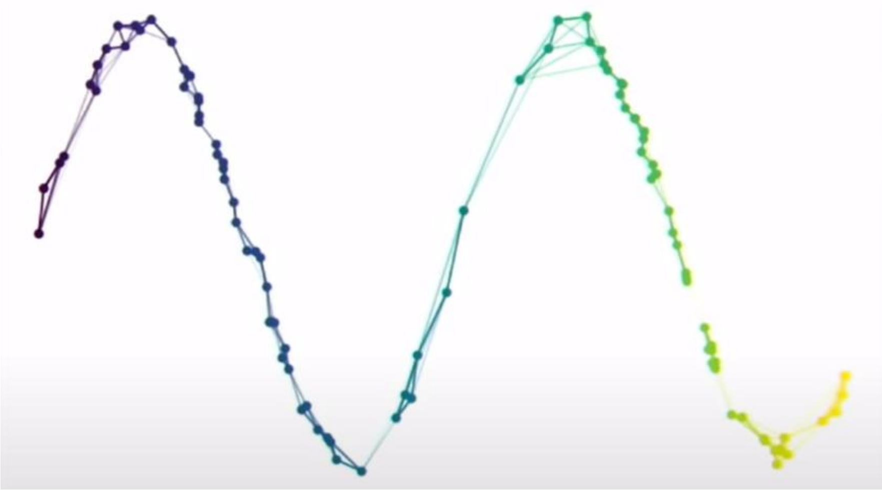




Source: “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction” McInnes et. al. 2018

Uniform Manifold Approximation and Projection



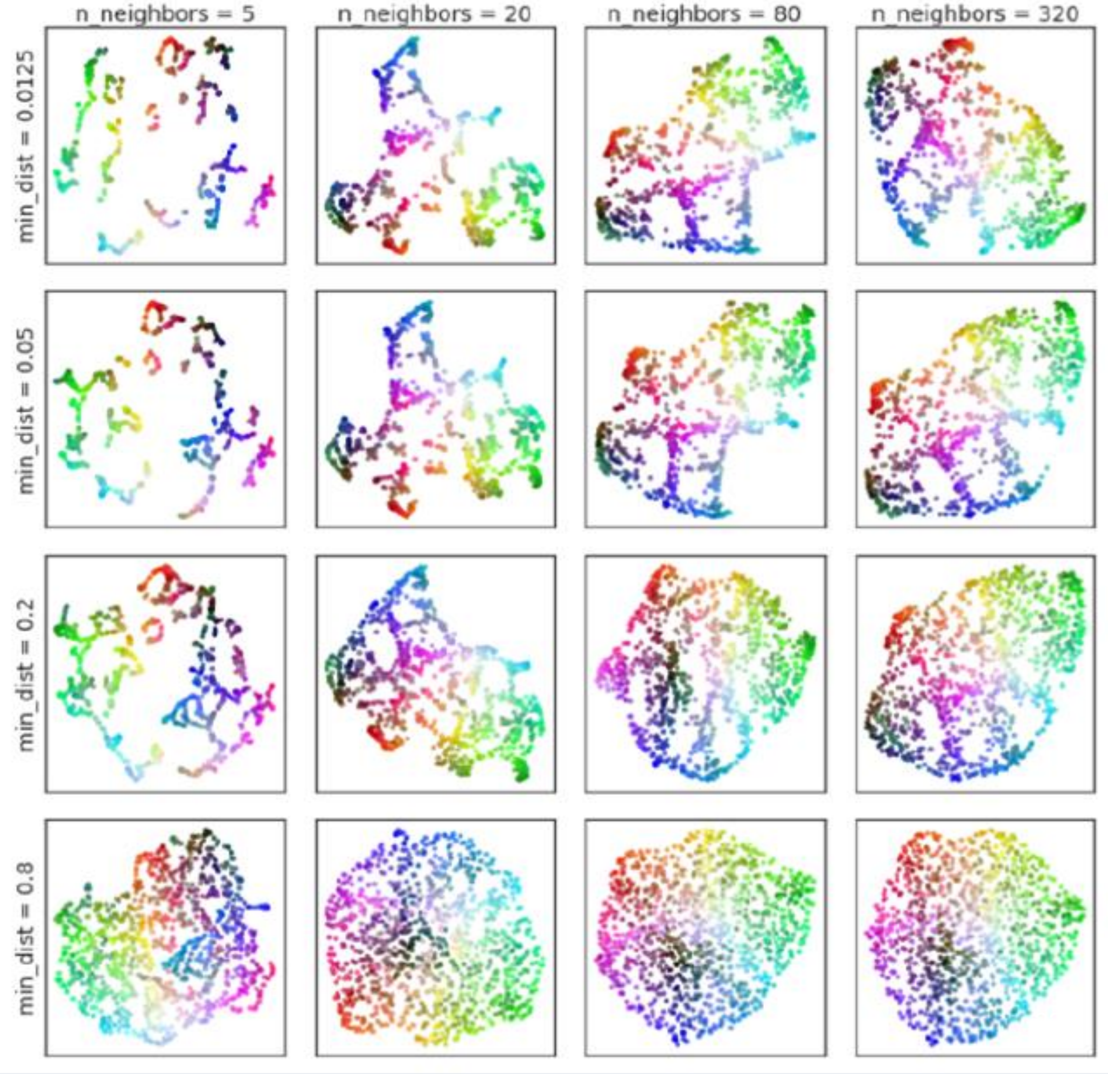


Source: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

UMAP parameters

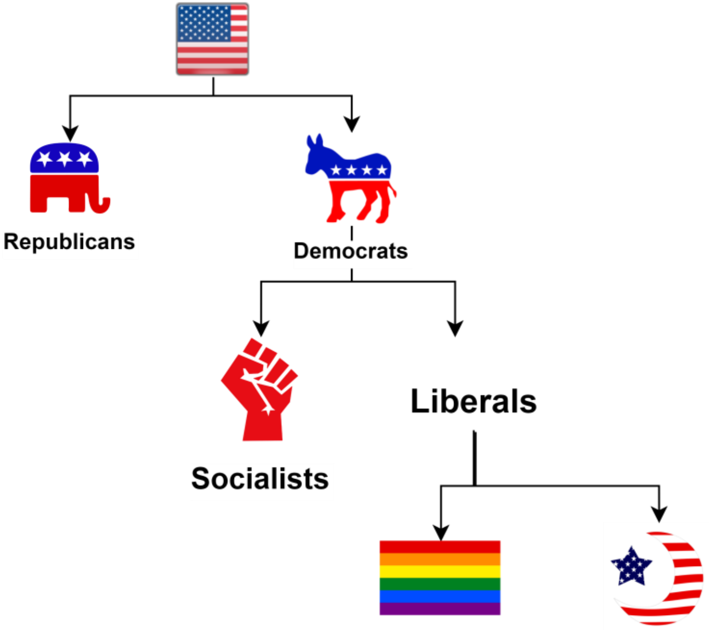
- **min_dist:** btw neighbors.
  - How tightly can UMAP pack elements together.
- **n_neighbors:** in a local structure.
  - Balance between local and global structures.

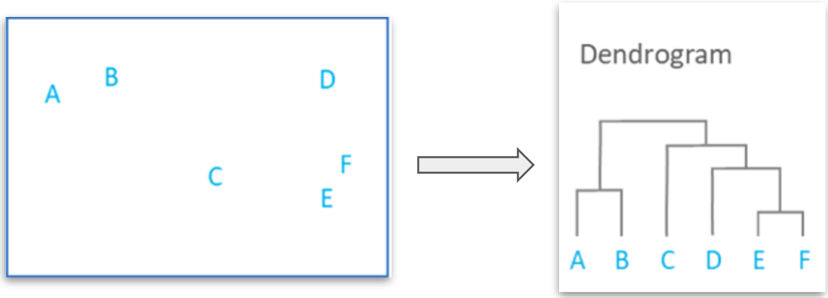
Source: “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction” McInnes et. al. 2018



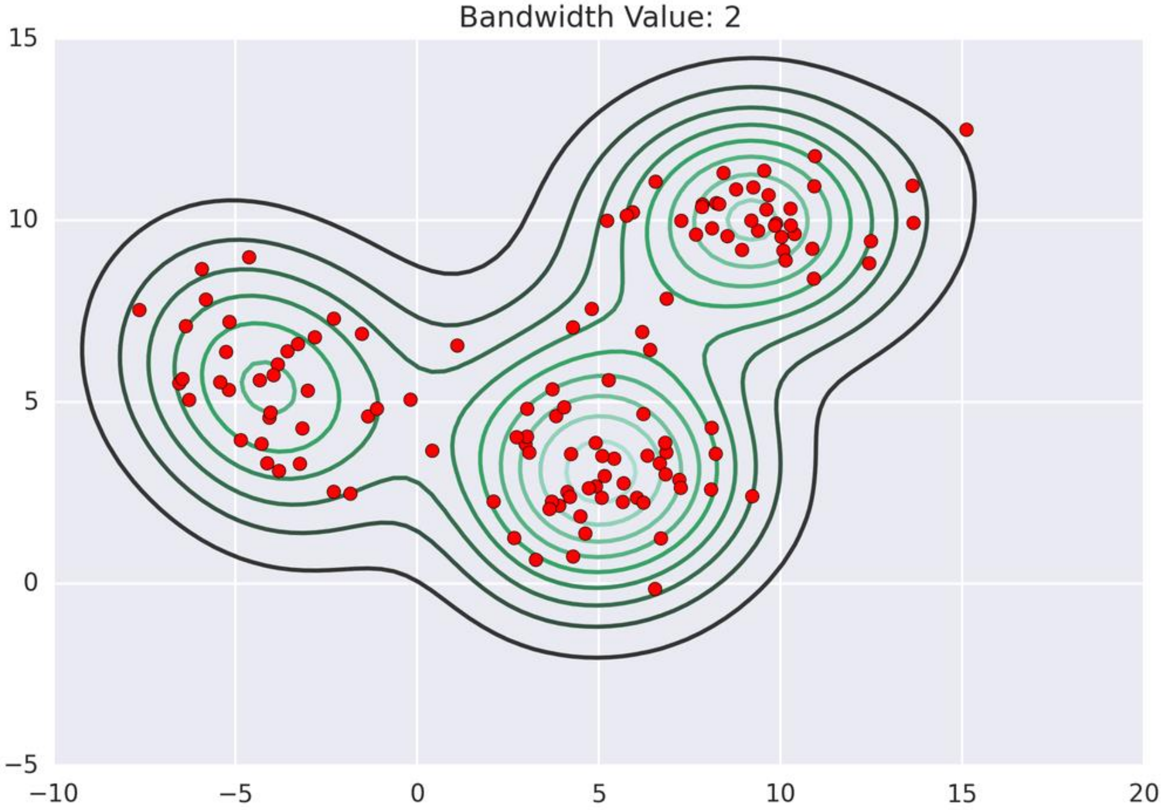
Clustering

	Flat	Hierarchical
Centroid/ Parametric	K-means GMM	Ward Complete-Linkage
Density/ Non-parametric	DBSCAN Mean-shift	HDBSCAN





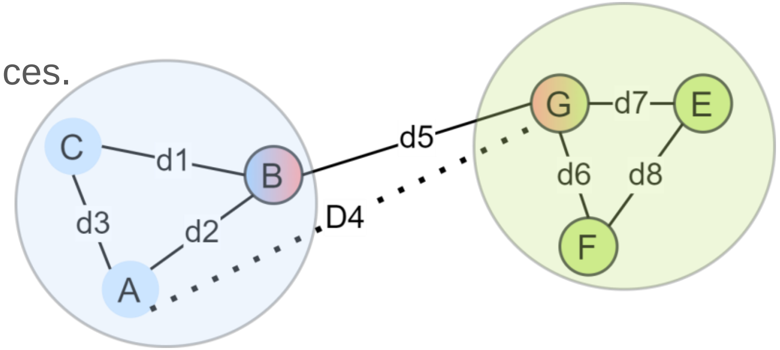
Meanshift



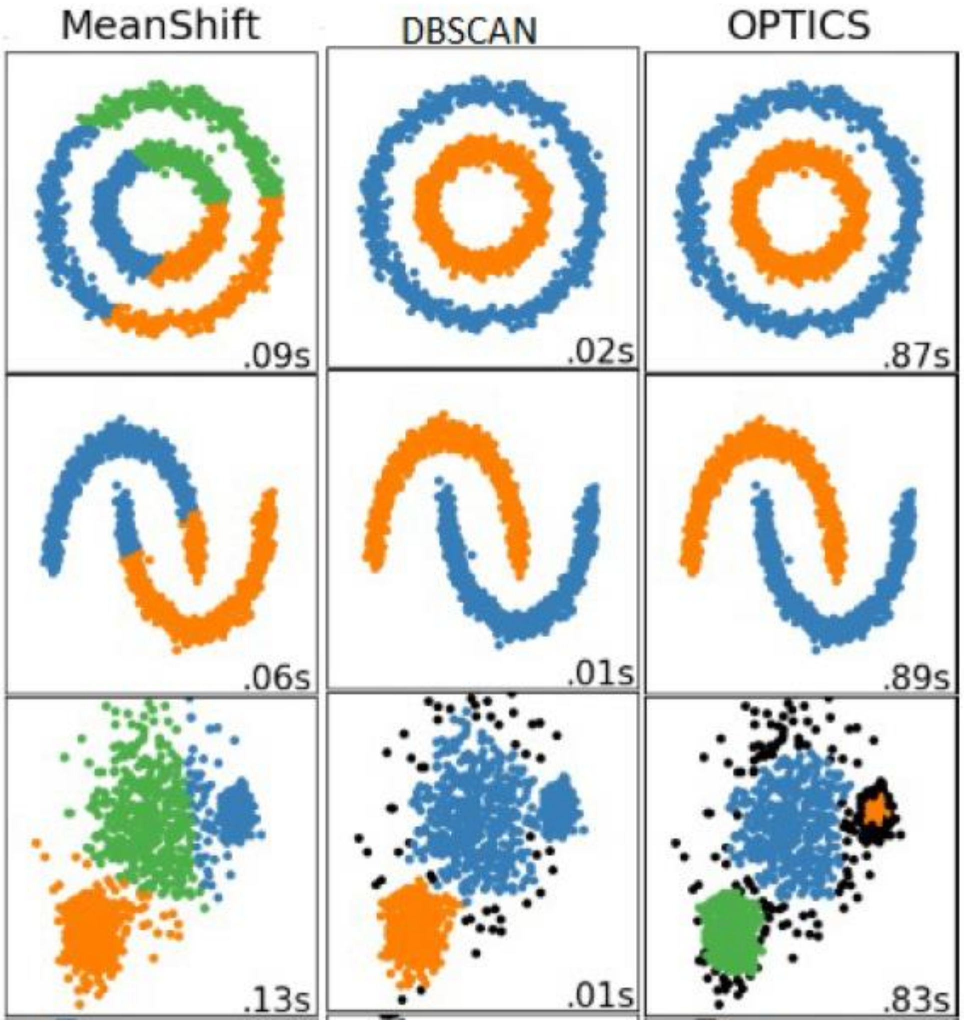
Source: https://informatics.sydney.edu.au/news/msg/

Graph-based modelling

- **Density-based spatial clustering of applications with noise (DBSCAN):**
  - **Edges:** if distance is at most ϵ.
  - **Core Object (CO):** a vertex connected to $\boldsymbol{min_{pts}}$ vertices.
  - **Density-connected:**
    - COs directly and transitively ϵ-reachable.
  - Weak in clustering data of varying densities.
  - It is hard to determine ϵ.

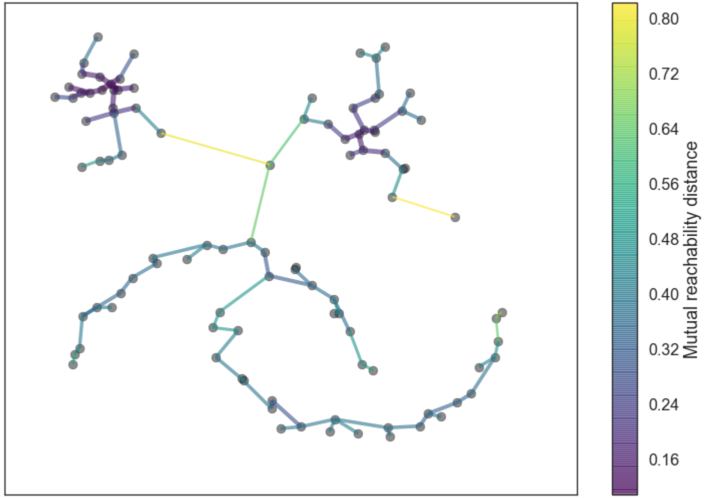


- **Ordering Points To Identify the Clustering Structure (OPTICS):**
  - **Core distance:** the distance to the nearest min_{pts} neighbors.
  - **Reachability distance:** $x \rightarrow y$ = max(dist (x, y), core distance of y)
  - Use priority-queues to represent density-based clustering structure

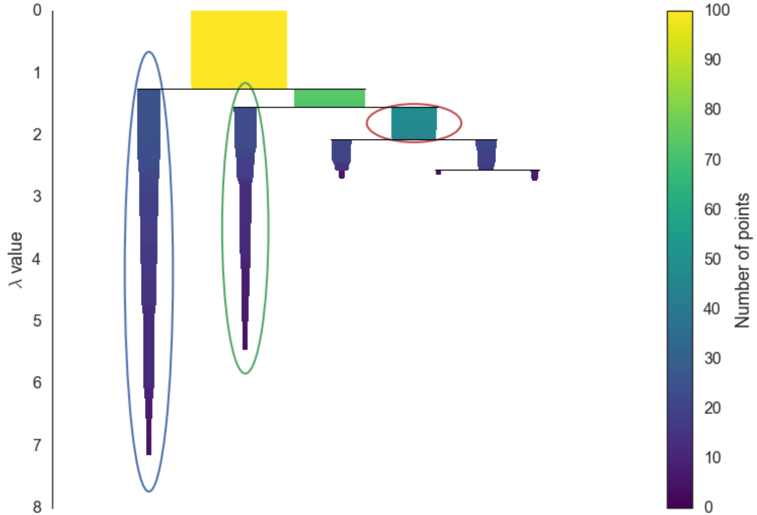


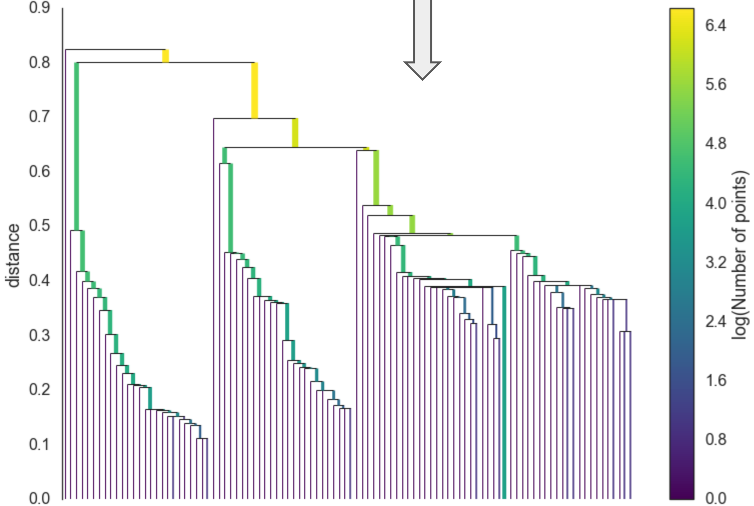
Source: https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

Hierarchical DBSCAN (HDBSCAN)



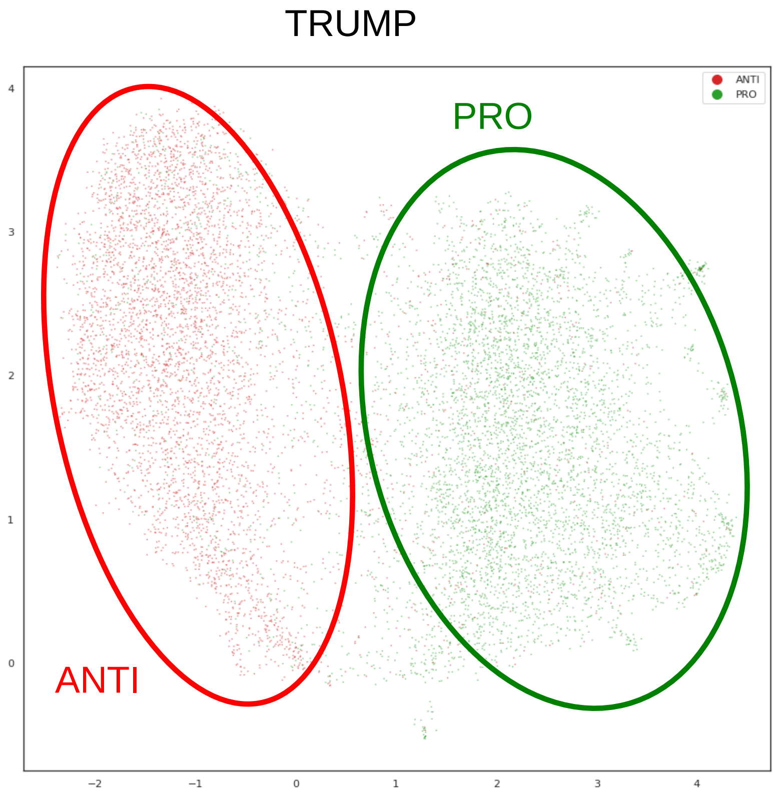
- Generates DBSCAN solutions for varying ϵ.
- Selects clusters with the best stability over ϵ.
- Simplify the tree using a minimum cluster size parameter.
- $\lambda = 1 / \text{distance}$

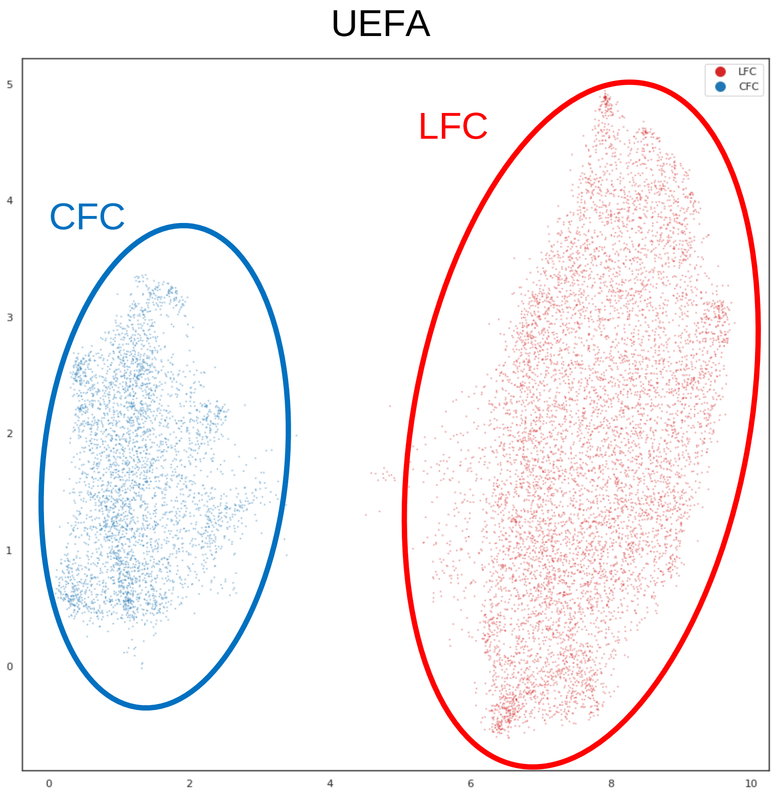


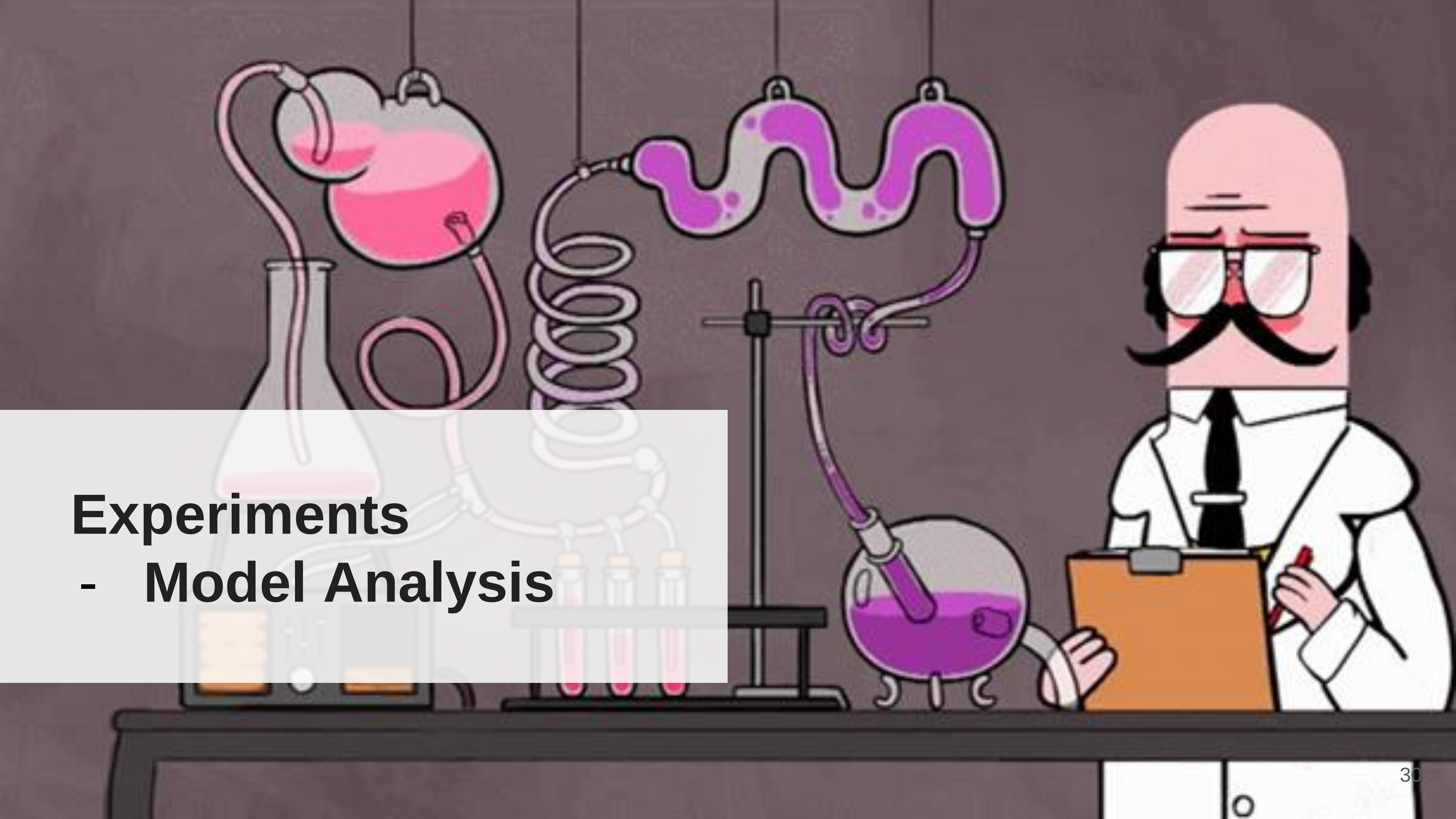


Source: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

Model

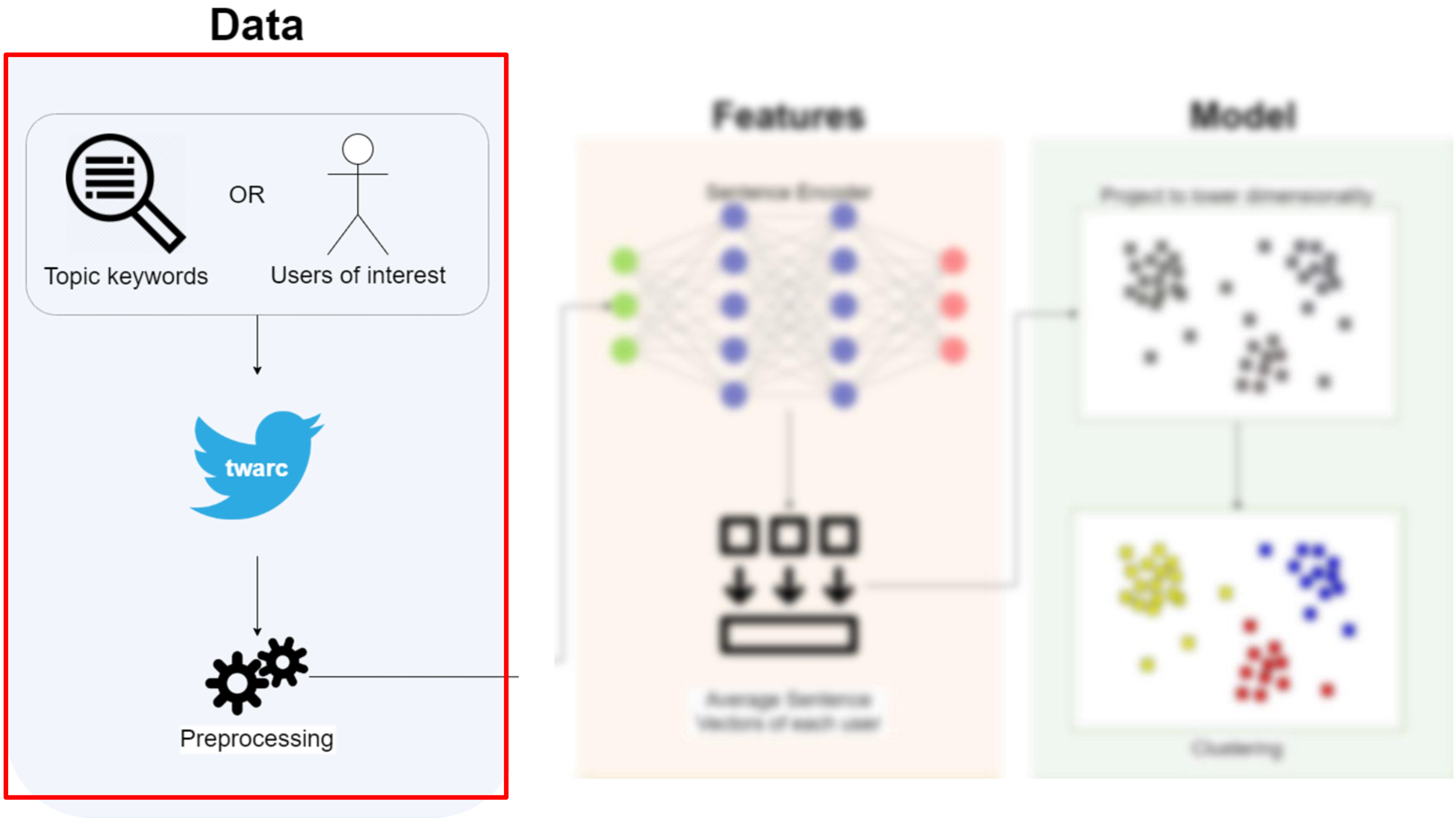




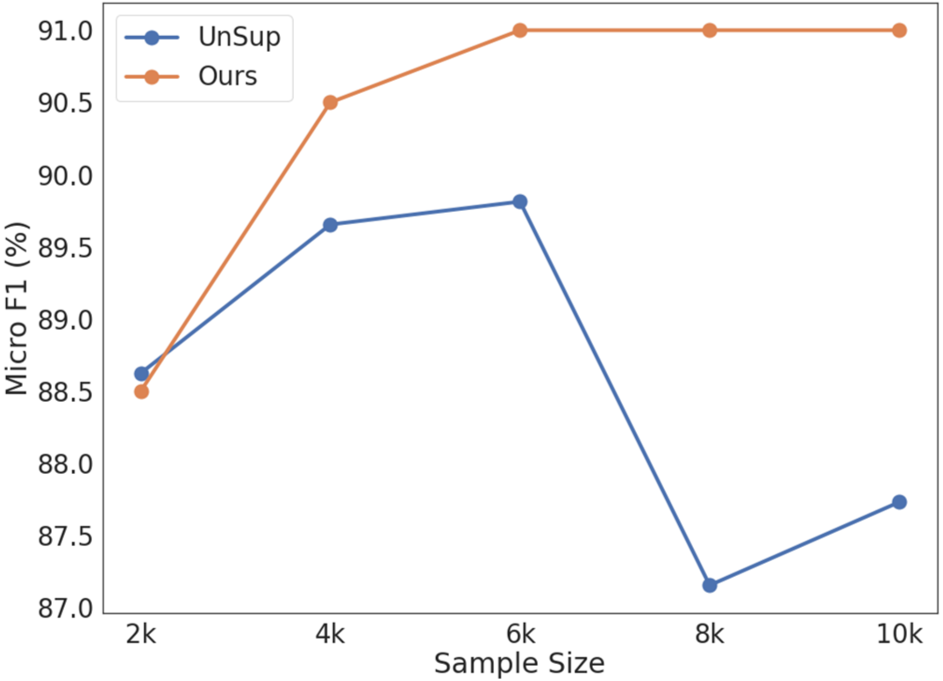


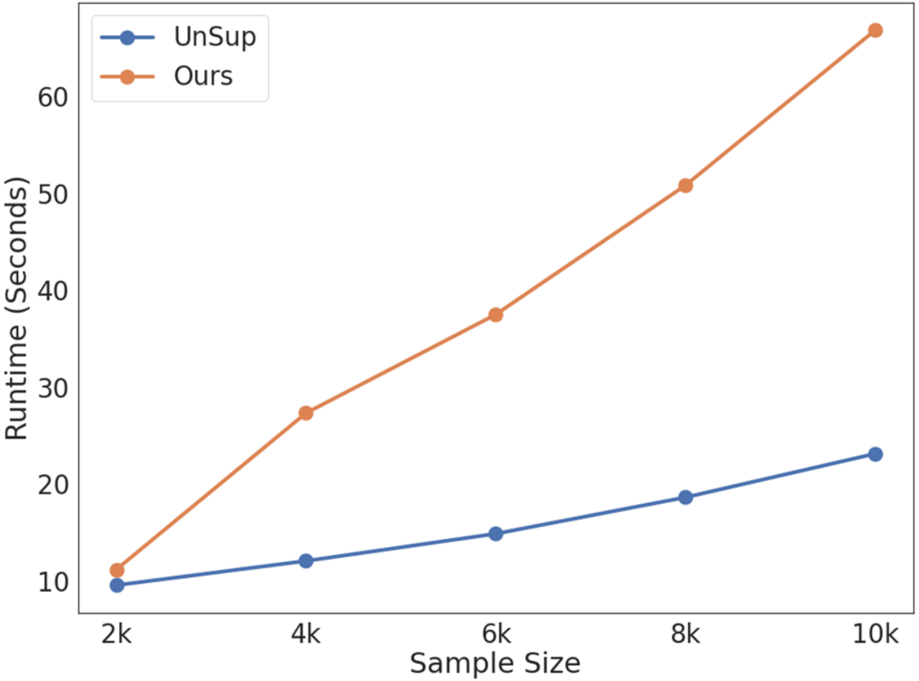
Experiments

- Model Analysis

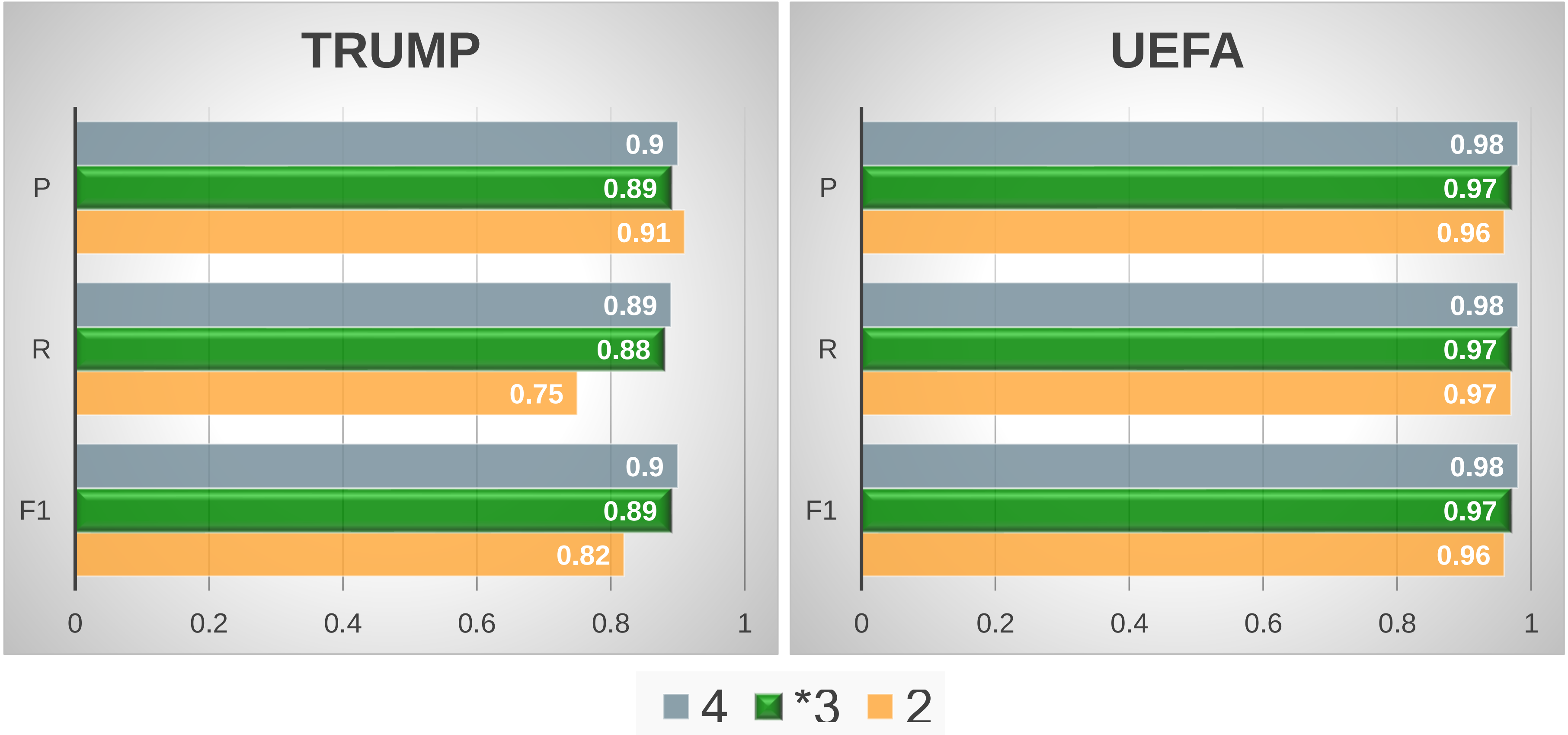

Data
Topic keywords
OR
Users of interest
twarc
Preprocessing
Features
Model

Data: Sample Size

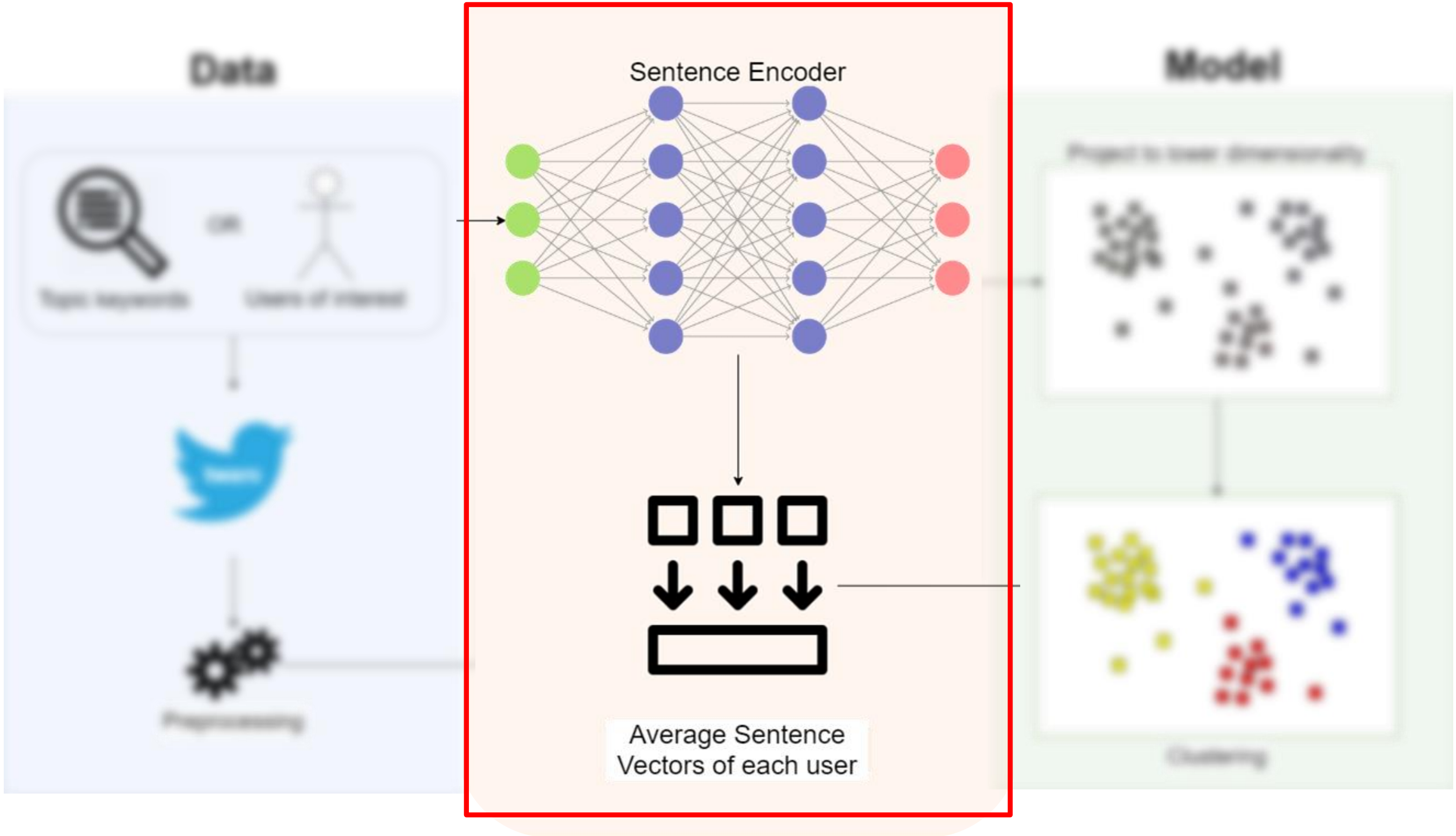




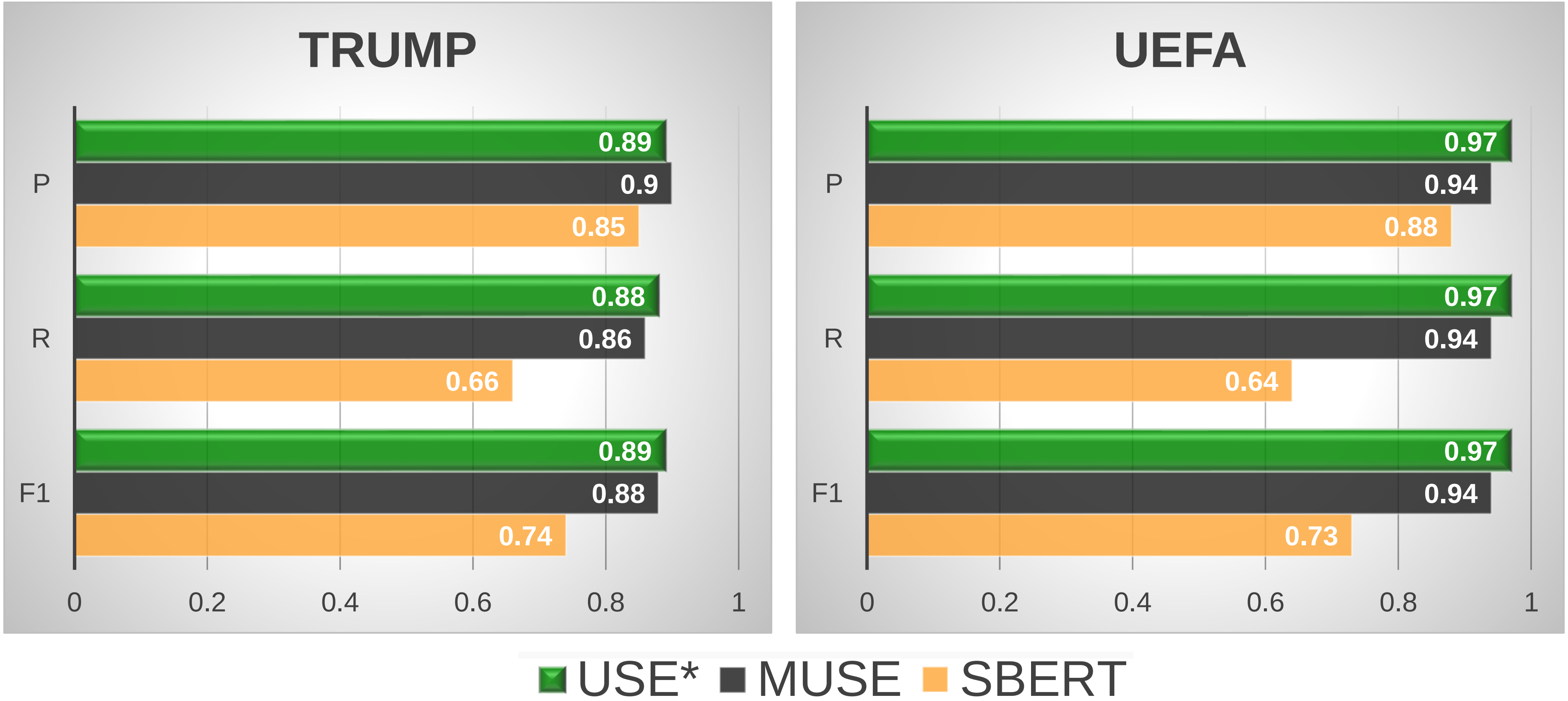
Data: Variance (min unique tweets per user)

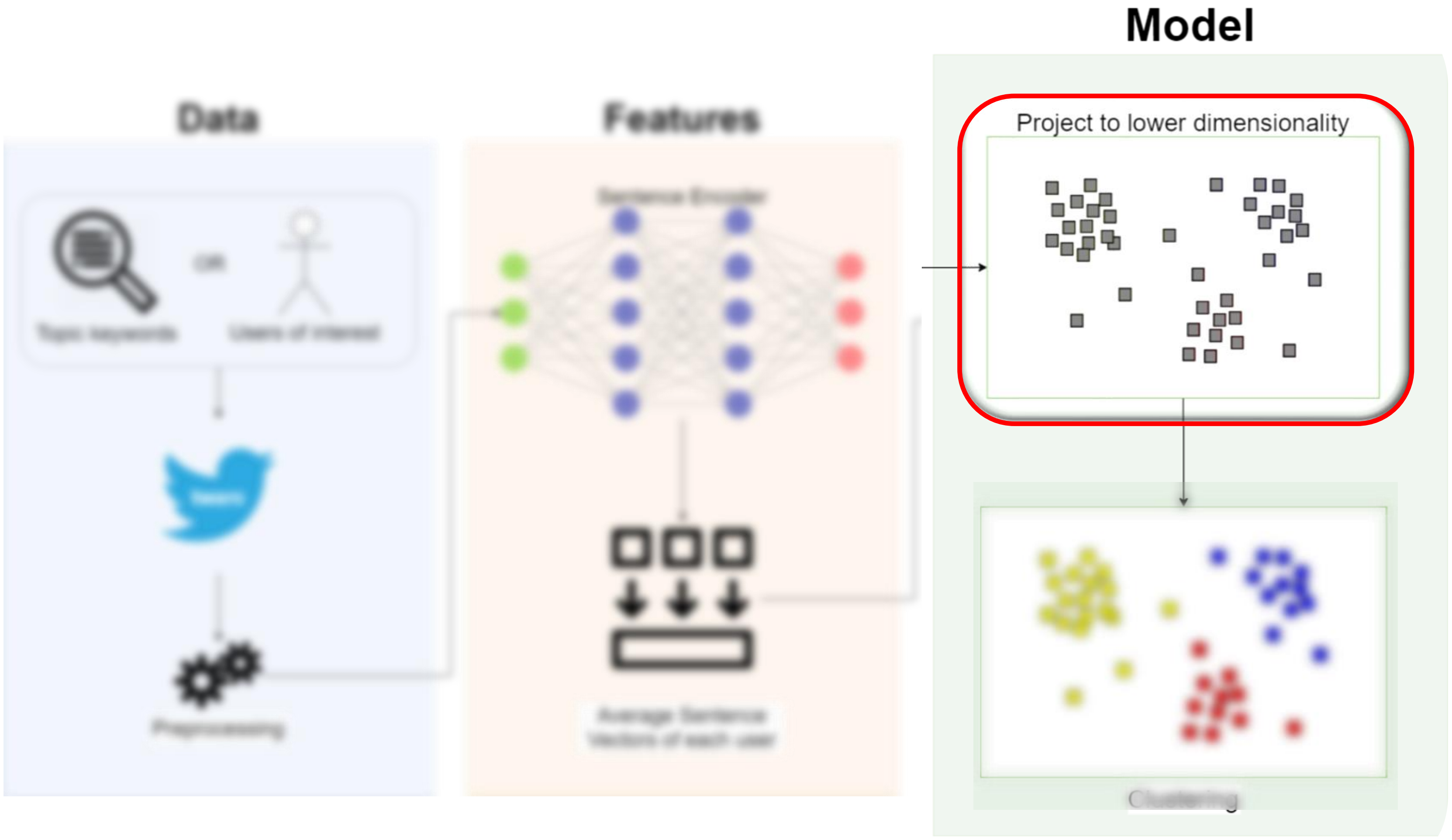


Features

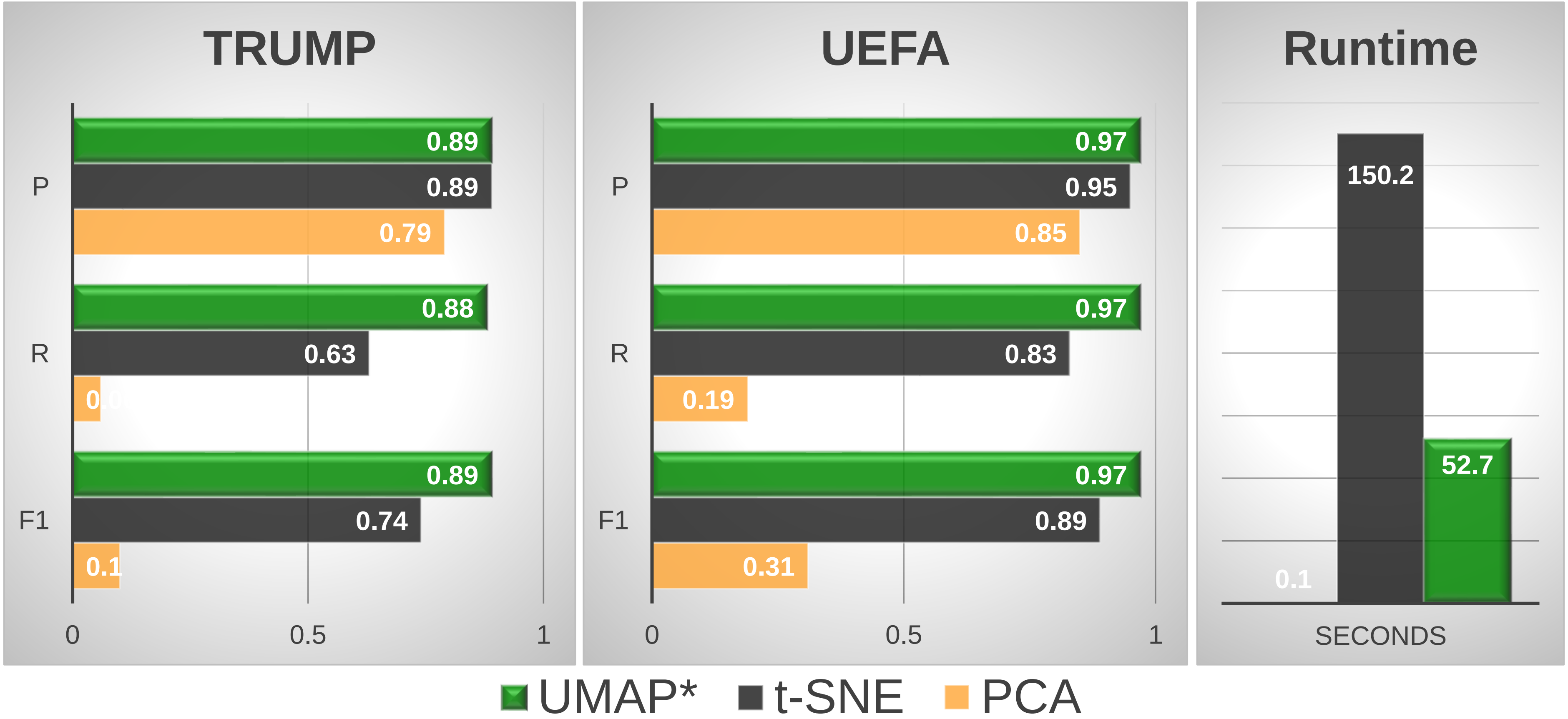

Data
Model
Sentence Encoder
Average Sentence
Vectors of each user

Features: Sentence Embeddings

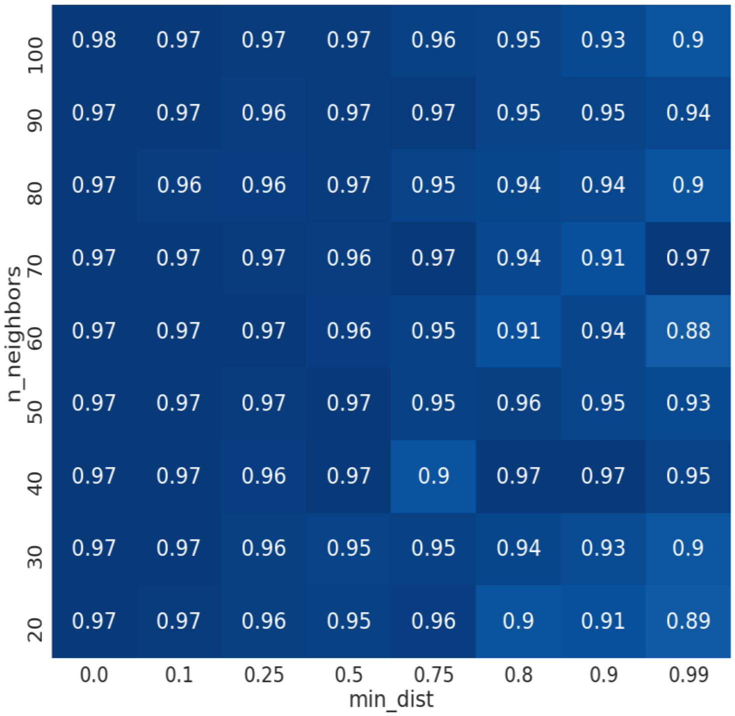



Model
Data
Features
Project to lower dimensionality

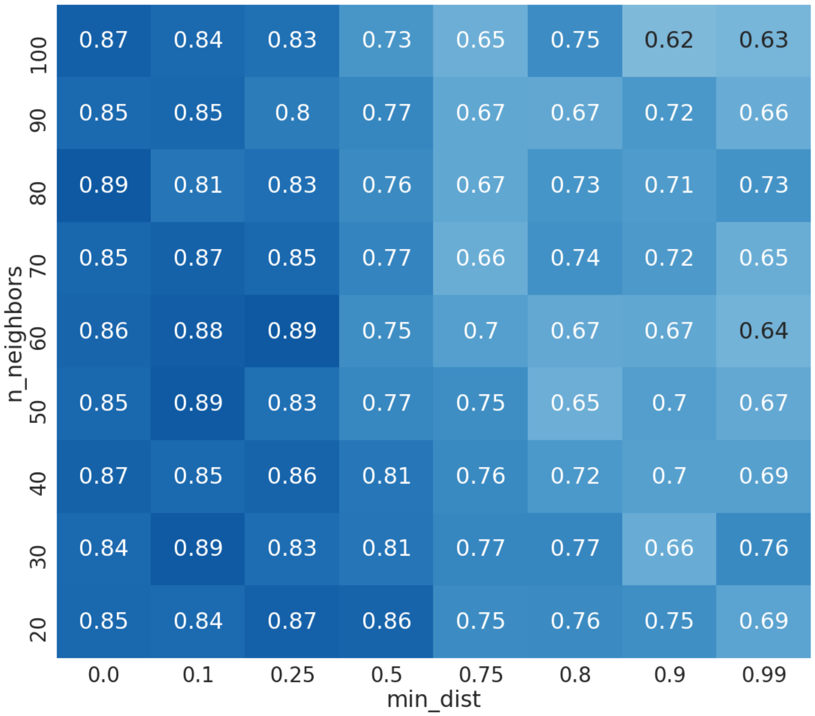
Model: Vector Projection



Model: Hyperparameters

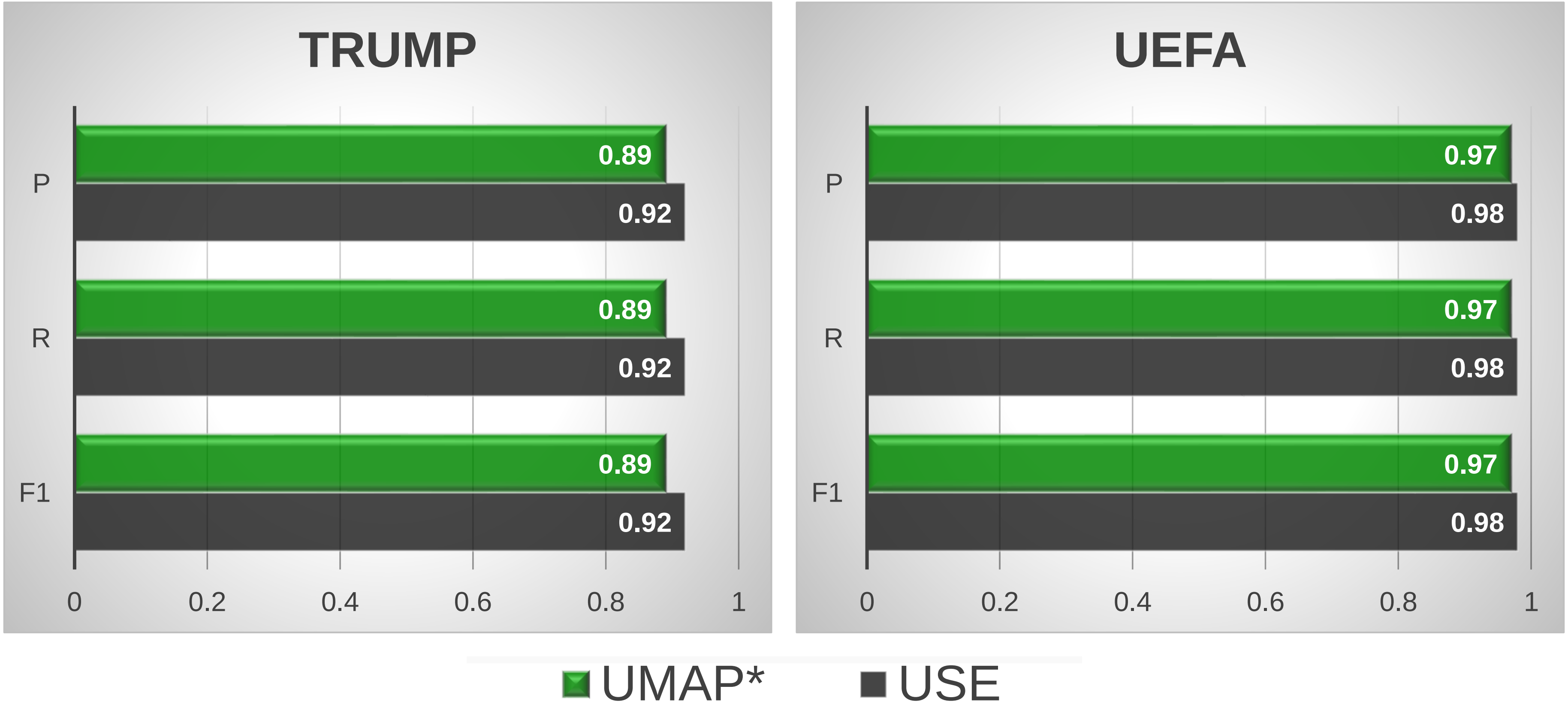


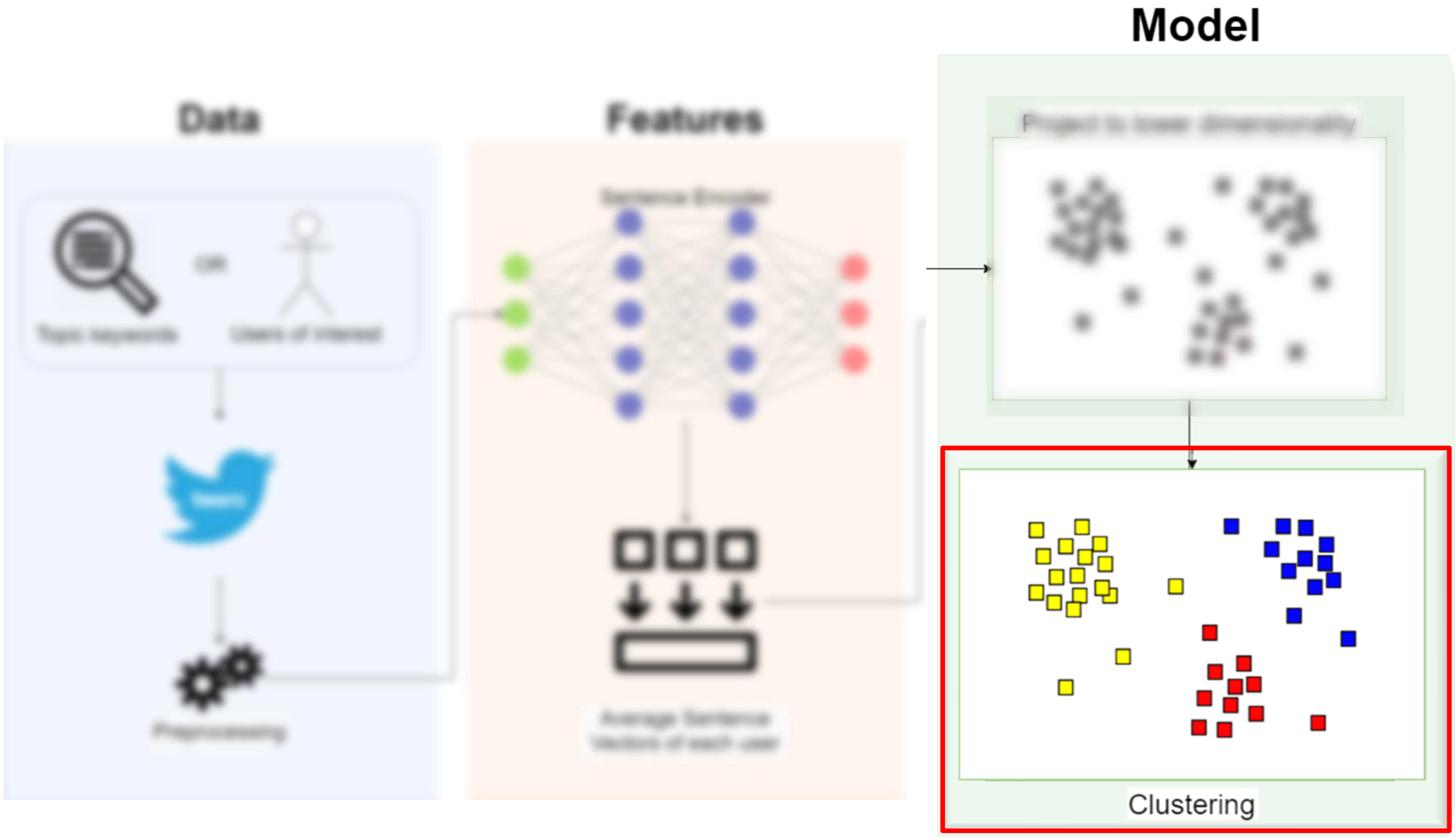
UEFA



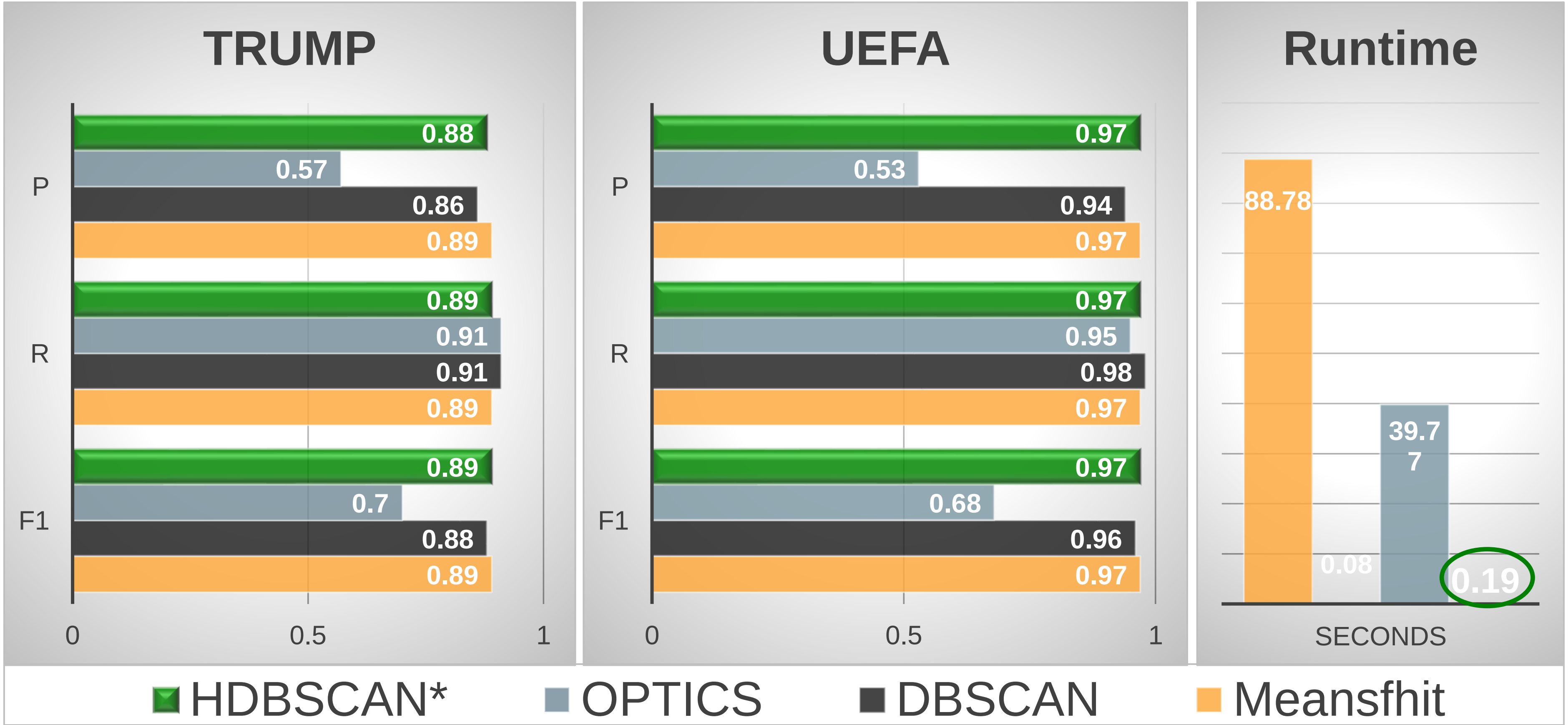
Trump

How much information is lost with UMAP?



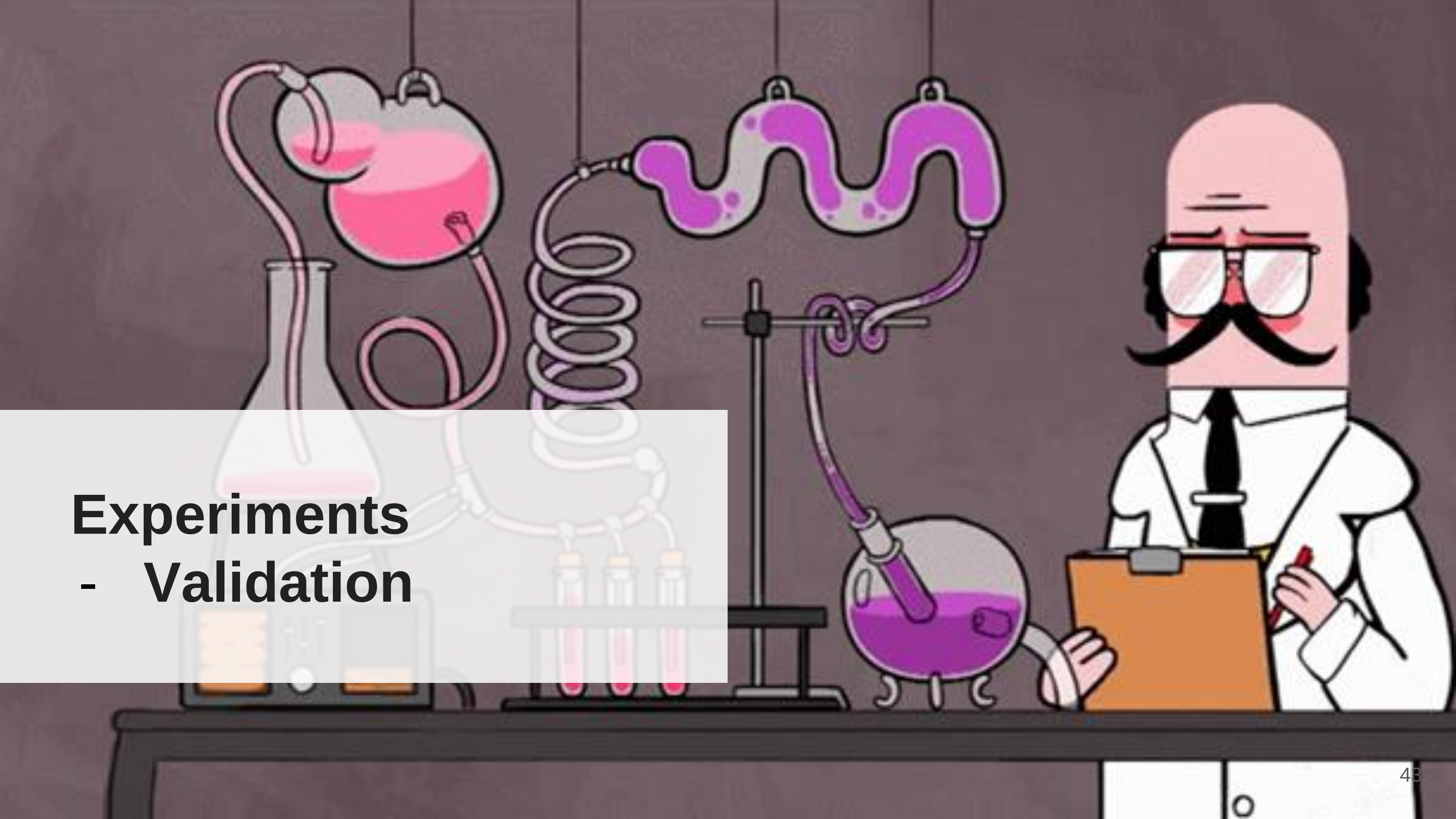

Model
Data
Features
Clustering

Model: Clustering



Limitations of previous best Unsupervised Stance Detection

- ➢ Requires platform-specific features. ✓
- ➢ Requires large amount of data for training. ✓
- ➢ Does not consider the hierarchy of stances. ✓
- ➢ Requires specification of the number of clusters. ✓



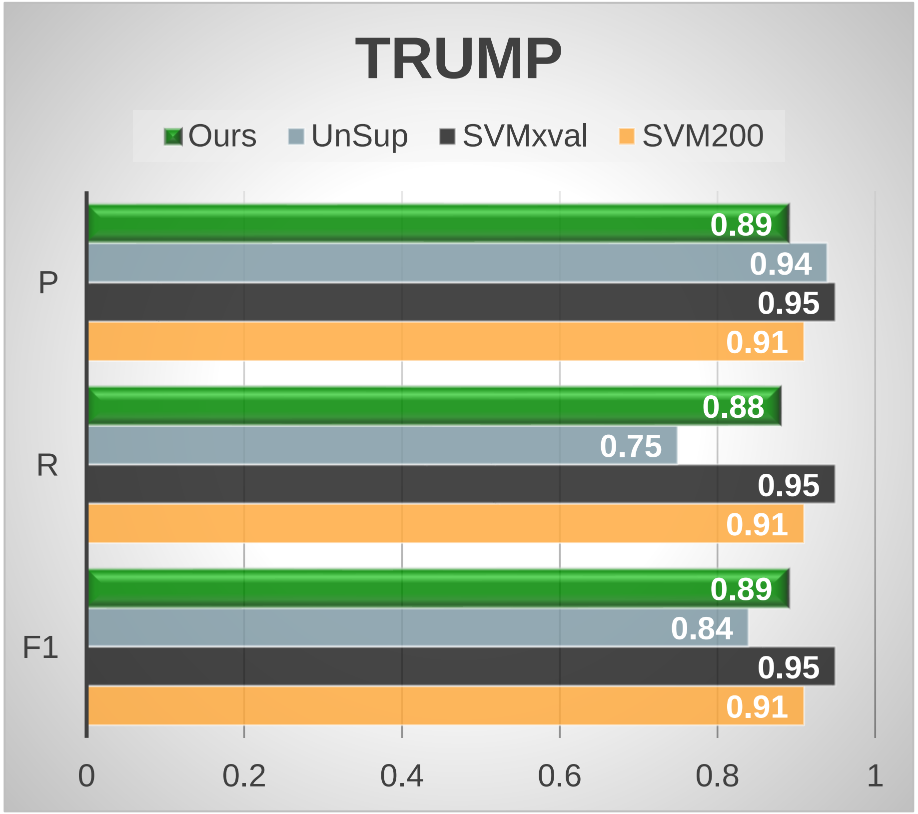
Experiments

- Validation

Baselines

- **Supervised:**
  - **Features:** retweeted accounts
  - **Models:**
    - **SVM$_{200}$:** Train on the most vocal 200 users in each label, test on the rest.
    - **SVM$_{xval}$:** 5-fold cross validation
- **Unsupervised**:
  - **Features:** retweeted accounts
  - **Model: UnSup**
    - Project user vectors of unique retweeted accounts using UMAP on a 2D plane
    - Cluster projected user vectors using Mean-shift

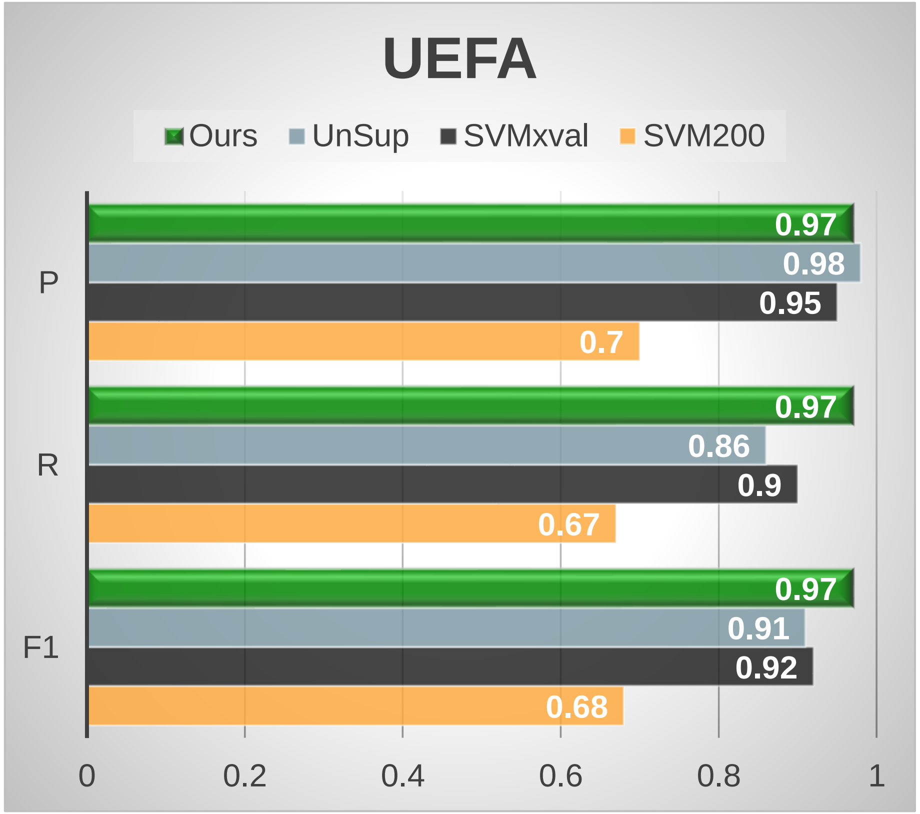
Results: Trump (full of #hashtags & retweets)



- Using just 200 users is enough for +90% precision.
- SVM: More samples = Better performance.
- UnSup fails in R, because of strong assumptions.
- Our approach maintains good R with high P.

Results: UEFA

- Hashtags and retweets behaviors are **NOT** led by just 200 users in a more distributed dataset.
- Hashtags and Retweets still show good results.
- UnSup maintains high P, but is still weak in R.
- Our approach excels in capturing semantic stance signals.



Case Study

- Quantitative Analysis

Topics

Topic	Keywords	Users	Tweets	Unique Tweets
Match	Super Cup, LivChe, CheLiv, Chelsea FC, Liverpool FC, YNWA, KTBFFH, LFC, CFC	42,095	336,373	92,687
Coaches	Lampard, Klopp	4,325	31,172	6,433
Stars	Salah, Mane, Firmino, Henderson, Van Dijk, Adrian, Kante, Tammy, Jorginho, Zouma, Pedro, Arrizabalaga	4,589	24,966	6,698
Referee	Referee, Frappart, VAR	1,170	2,806	774
	Total	52,179	395,317	106,592

Can we infer the stance on a topic from a different topic?

Stance Alignment: Method

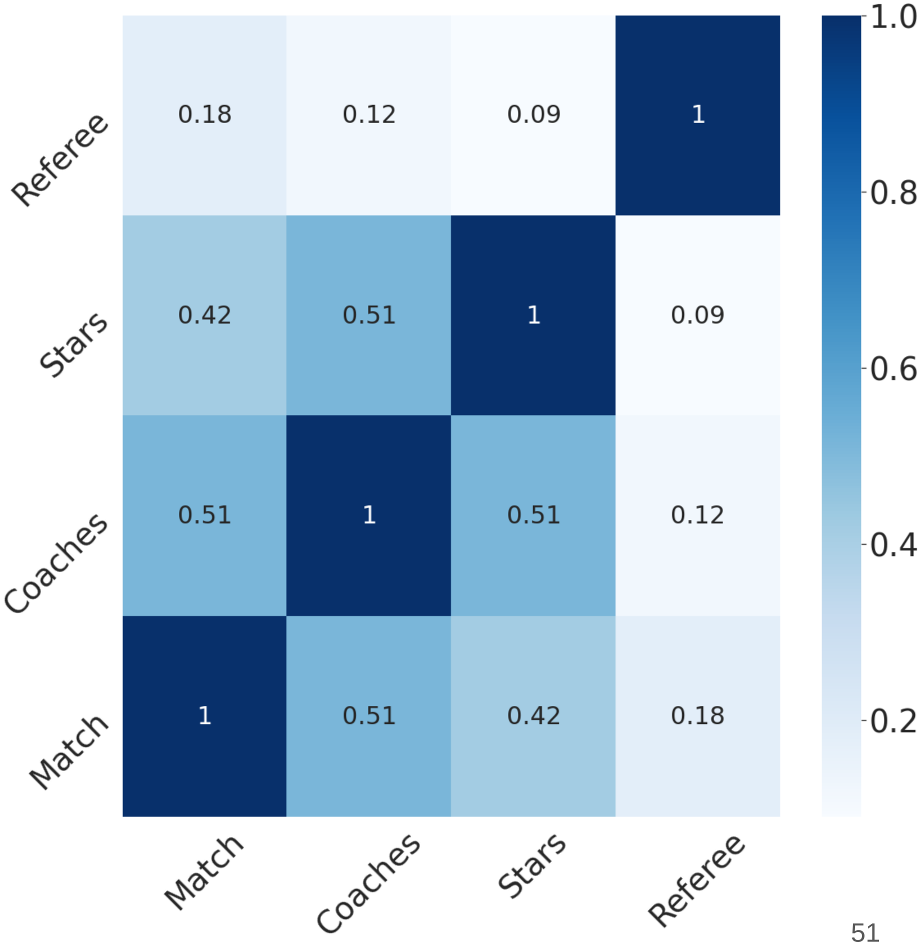
- Normalized mutual information adjusted for randomness (AMI).
- Measures mutual information, or mutual independence, between two clustering solutions.

$$MI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} log \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

$$AMI(U,V) = \frac{MI(U,V) - E(MI(U,V))}{avg(H(U),H(V)) - E(MI(U,V))}$$

Stance Alignment: Results

- **Similar**:
  - *Match*, *Coaches*, *Stars*.
- **Different:**
  - Referee
- stances towards clubs are dependent on the stances towards the coaches and the stars, rather than the referee



What does the polarization degree inform us on different topics?

Polarization Quantification: Method

- Random Walk Controversy (RWC).
- RWC computes the shortest graph traversal from a set of random users to prominent users either with the same or different stances.
- prominent nodes are the top n nodes with the most connections in a community.

$$RWC = P_{AA}P_{BB} - P_{AB}P_{BA}$$

- Where
    - A and B are different Classes.
    - P_{XY} is the probability that a random node in X would reach highly connected node in Y.

Polarization Quantification: Result

- **Higher** polarization on *Coaches*
  - Each cluster is talking mainly about their own coach.
- **Moderate** polarization on *Stars*
  - Users from both groups talk about the stars of other teams certain level of semantic similarity.
- **Lower** polarization on *Referee*
  - Can be attributed to the small size of the dataset

Topic	**RWC**
Match	0.62
Coaches	0.95
Stars	0.78
Referee	0.73

Case Study
- Qualitative Analysis

Semantic Difference in Clusters

- Valence x log Frequency = Prominence
- Prominence of a term t in a set of tweets a, as opposed to the set b

$$\Pr(t, D_a, D_b) = \log(tf_{t,D_a}) \times \left(2\frac{\frac{tf_{t,D_a}}{|D_a|}}{\frac{tf_{t,D_a}}{|D_a|} + \frac{tf_{t,D_b}}{|D_b|}} - 1\right)$$

- ***t***: term
- $\boldsymbol{D_x}$: set of tweets x
- $\boldsymbol{tf_{t,Dx}}$: term frequency of the term t in the set of tweets x
- $|\boldsymbol{D_x}|$: sum of the frequencies of all terms in x

(a) Match (CFC) (b) Match (LFC) (c) Coaches (CFC) (d) Coaches (LFC)

(e) Stars (CFC) (f) Stars (LFC) (g) Referee (CFC) (h) Referee (LFC)

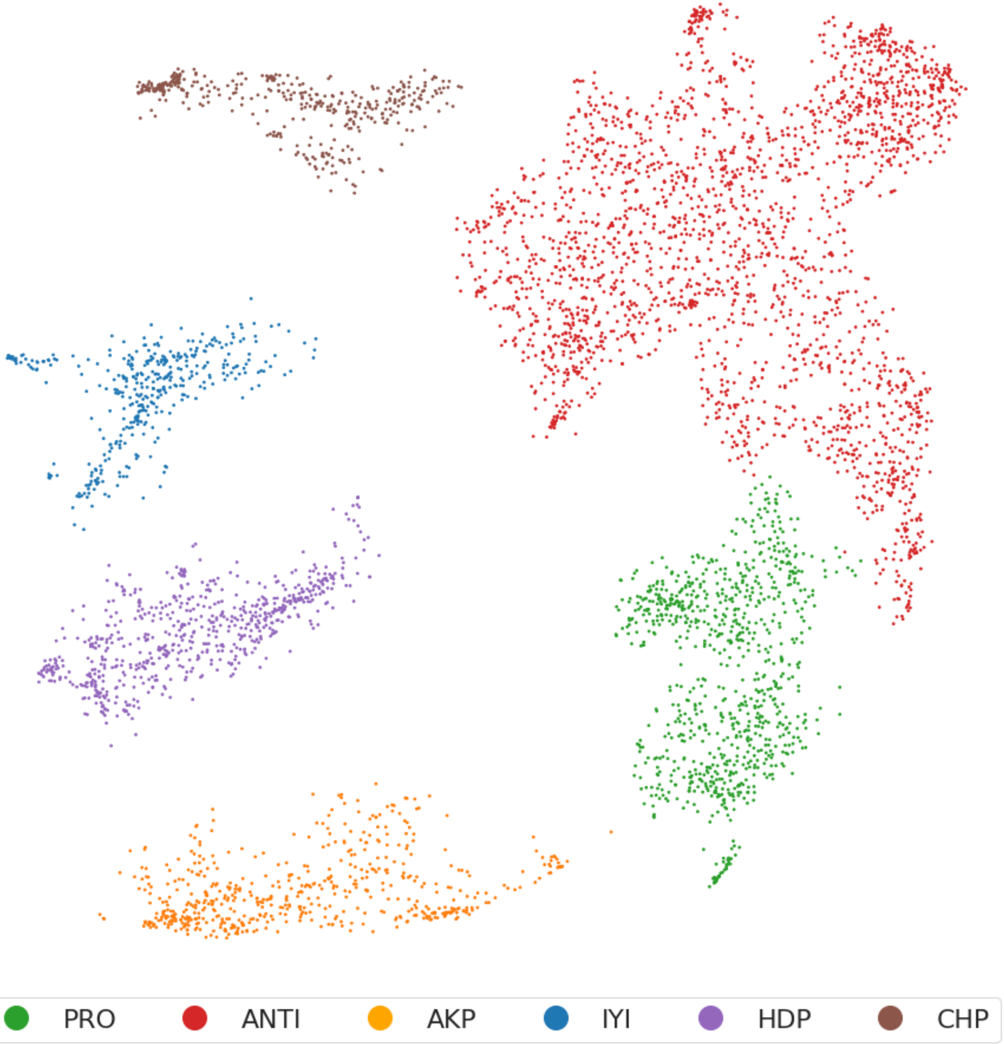
Word Clouds Generated by prominent terms in each cluster for different topics.

Limitation Analysis

- **Selective retweeting:**
  - RT @LFC: Firmino scores the first penalty of the shootout!
- **Cross Mentioning:**
  - I respect Liverpool and they had chances in the game but it was clear to me that we were the stronger team
- **Implicit Criticism:**
  - We won the Super cup with the worst performance since Klopp took over haha #LIVCHE
- **Post-event Teasing:**
  - @Calabastu VAR knows your penalty wasn't the right call #lfc #cfc
- **Pre-event Teasing:**
  - #Lampard it’s only #Liverpool on Wednesday #cfc #lfc #ynwa

ICWSM: TR elections

	Arab	CHP	Erdoğan	HDP	PKK	Syrian	Trump	USA
USA	0.89	0.94	0.82	0.95	0.81	0.88	0.95	1
Trump	0.9	0.92	0.8	0.93	0.86	0.86	1	0.95
Syrian	0.87	0.87	0.74	0.87	0.76	1	0.86	0.88
PKK	0.81	0.81	0.7	0.83	1	0.76	0.86	0.81
HDP	0.9	0.93	0.79	1	0.83	0.87	0.93	0.95
Erdoğan	0.75	0.78	1	0.79	0.7	0.74	0.8	0.82
CHP	0.89	1	0.78	0.93	0.81	0.87	0.92	0.94
Arab	1	0.89	0.75	0.9	0.81	0.87	0.9	0.89



Summary
&
Conclusion

Summary

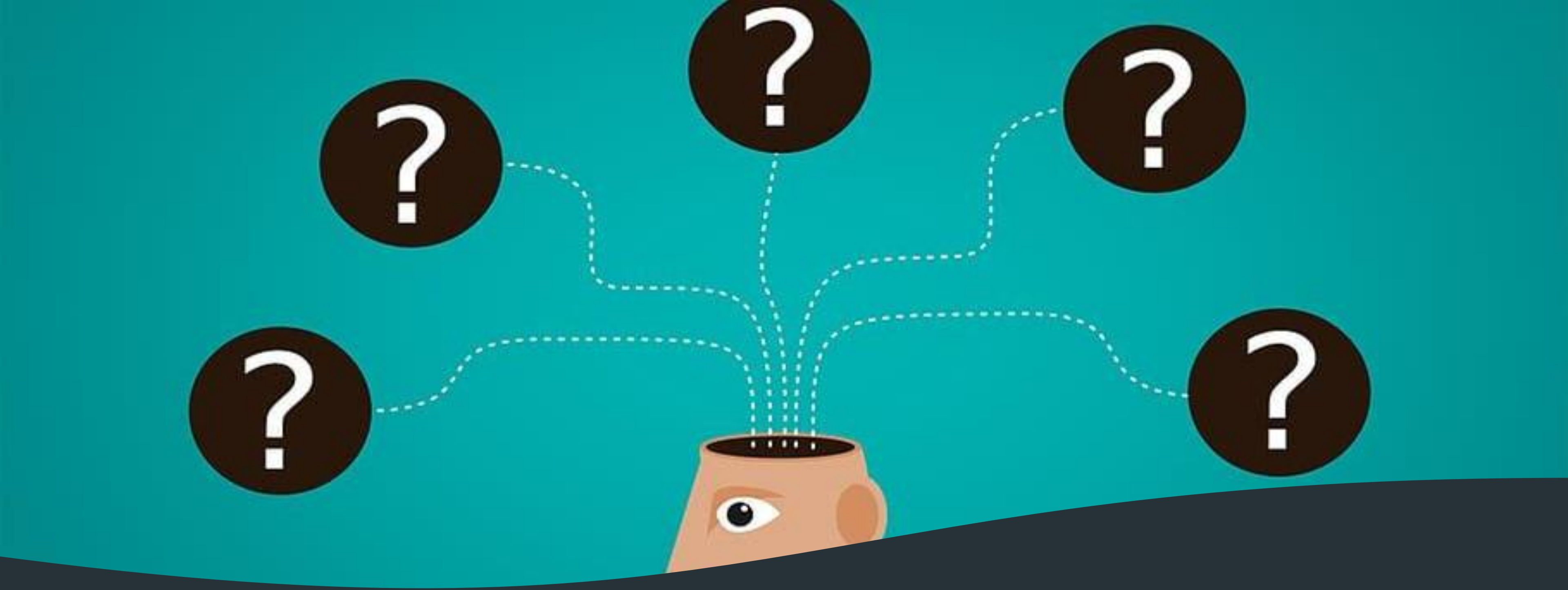
- **Methodology:**
  - Collect tweets using keywords or usernames.
  - Encode tweets using a pre-trained sentence encoder, e.g. Universal Sentence Encoder.
  - Average tweet vectors per user.
  - Project user vector to a lower dimensional space, e.g. using UMAP.
  - Cluster projected vectors, e.g. using HDBSCAN.
- **Analysis:**
  - Stance Alignment with AMI.
  - Polarization Quantification with RWC.
  - Semantic Difference using prominence score.

Contributions & Novelty

- **Performance:**
  - Competitive with supervised approaches.
  - Outperform previous unsupervised approaches.
- **Requirements:**
  - No platform-specific features required.
  - No fine-tuning needed.
- **Features:**
  - First to use pre-trained sentence embeddings for stance detection.
  - First to provide hierarchical clustering solution for stances,
- **Study:**
  - Compiled a dataset of +300k tweets and +12k stance-labeled users.
  - Provide quantitative and qualitative analysis of polarization in sports.

While at Özyeğin

- **Full-paper**: accepted for ICWSM'21
  - Embeddings-Based Clustering for Target Specific Stances: The Case of a Polarized Turkey.
- **Workshop**: Ranked 1st in the shared task, LREC | OSACT (May. 2020)
  - ALT Submission for OSACT Shared Task on Offensive Language Detection.
- **Poster**: Presented at SocInfo (Nov. 2019)
  - Embedding-based Qualitative Analysis of Polarization in Turkey.
- **Under review:** Submitted at COLING'20
  - Arabic offensive language on twitter: Analysis and experiments. (already +13 citations)



Questions