SDS2021

SWISS CONFERENCE ON DATA SCIENCE

UNI FR

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

GraphEDM: A Graph Based Approach for Entity Disambiguation in Microposts

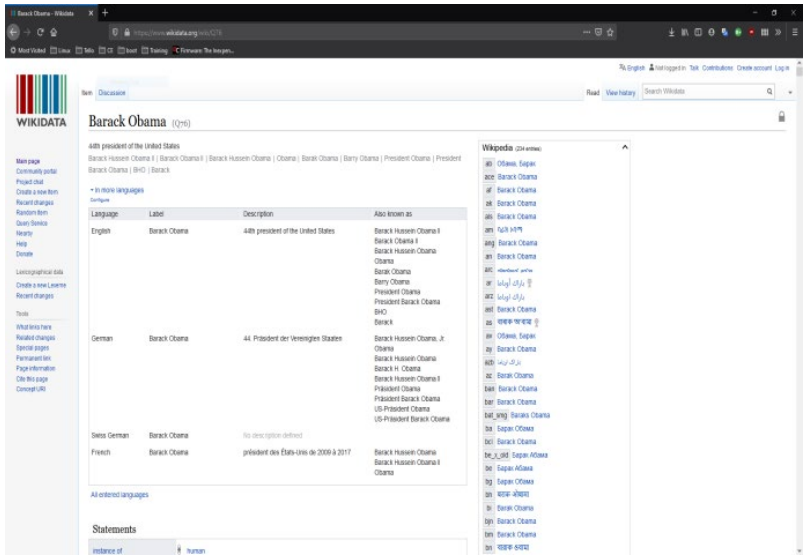
Prathyusha Nerella, Akansha Bhardwaj, Paolo Rosso, Philippe Cudré-Mauroux

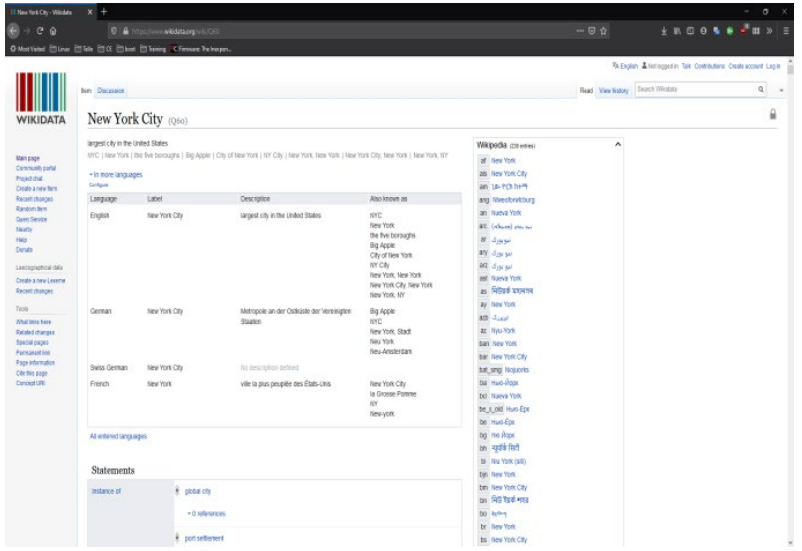
09 June, 2021

Table of Contents

Introduction

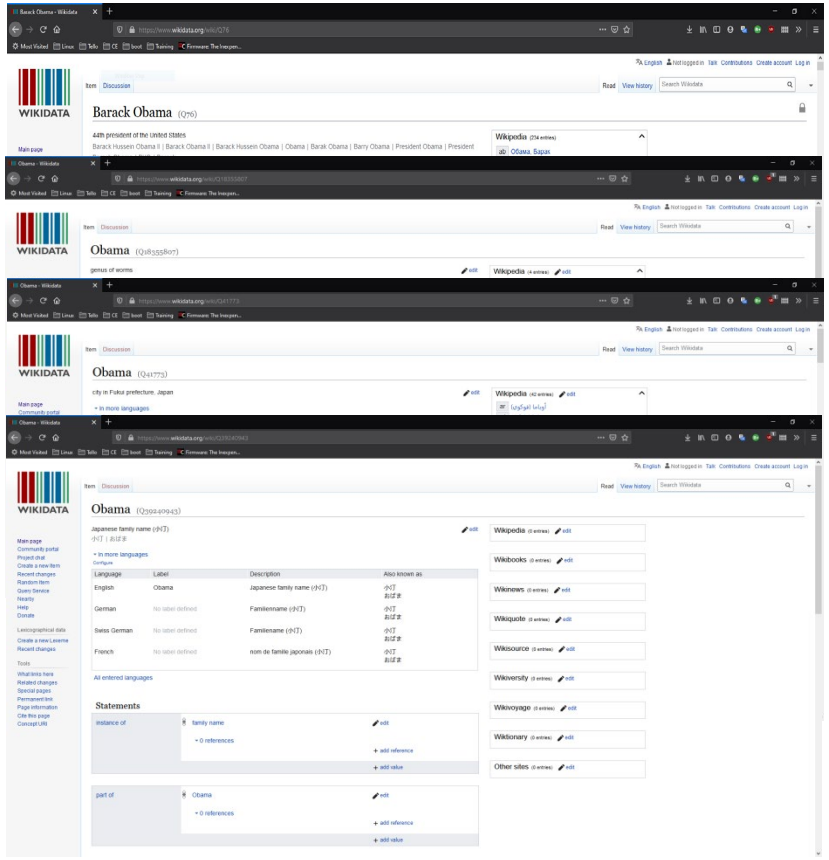
- **Entity Disambiguation:** a sub-field of **IE** that maps **mentions** in text to **entities** in a reference **Knowledge Base**
- Example: Obama visited Lake View Restaurant in New York City yesterday
- Entities from Wikidata Knowledge Base





Need for Disambiguation

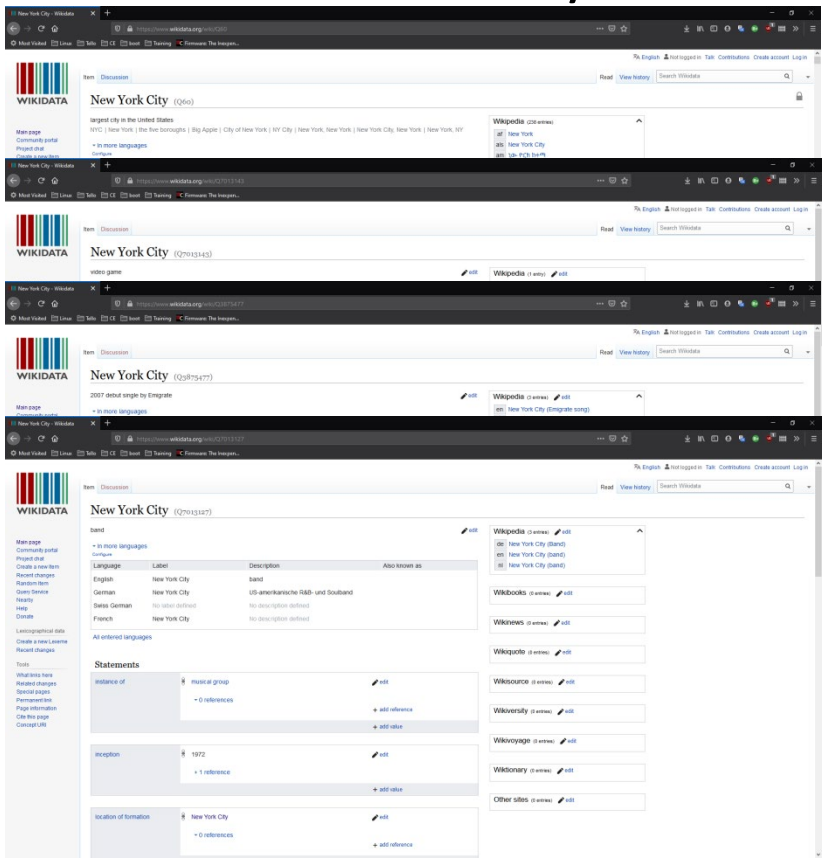
Obama



....
....

9 possible entities

New York City



....
....

11 possible entities

Named Entity Disambiguation in Microposts

- **Twitter** is one of the widely used micro blogging services
- About 500 Million tweets are posted per day
- Tweets are valuable source of information for different tasks
- Semantic understanding of tweets is required for tasks involving tweet analysis
- Entity disambiguation helps in understanding tweets semantically

Challenges for Entity Disambiguation in Microposts

- **Limited length:** character limit makes users tweet concisely making it difficult for automatic entity disambiguation

- **Informal style:** emojis, abbreviations, hashtags, missing punctuation or capitalization make interpretation difficult

Related Work

- Different techniques have been proposed for Entity Disambiguation on documents, tables and also on informal short texts

- Named Entity Extraction and Linking (NEEL) methods focus on both entity extraction and disambiguation tasks

- Wikidata, YAGO, DBpedia Knowledge Bases

Methodology

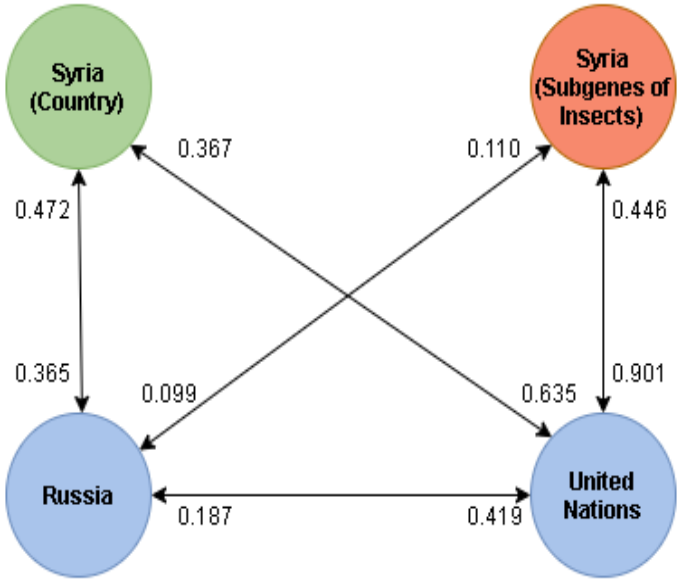
- GraphEDM approach is divided into two phases
- **Context Extension phase:** enhancement of tweet context to aid entity disambiguation process
- **Entity Disambiguation phase:** disambiguation of entities with a graph based approach leveraging the extended context
- **Wikidata** Knowledge Base

Context Extension

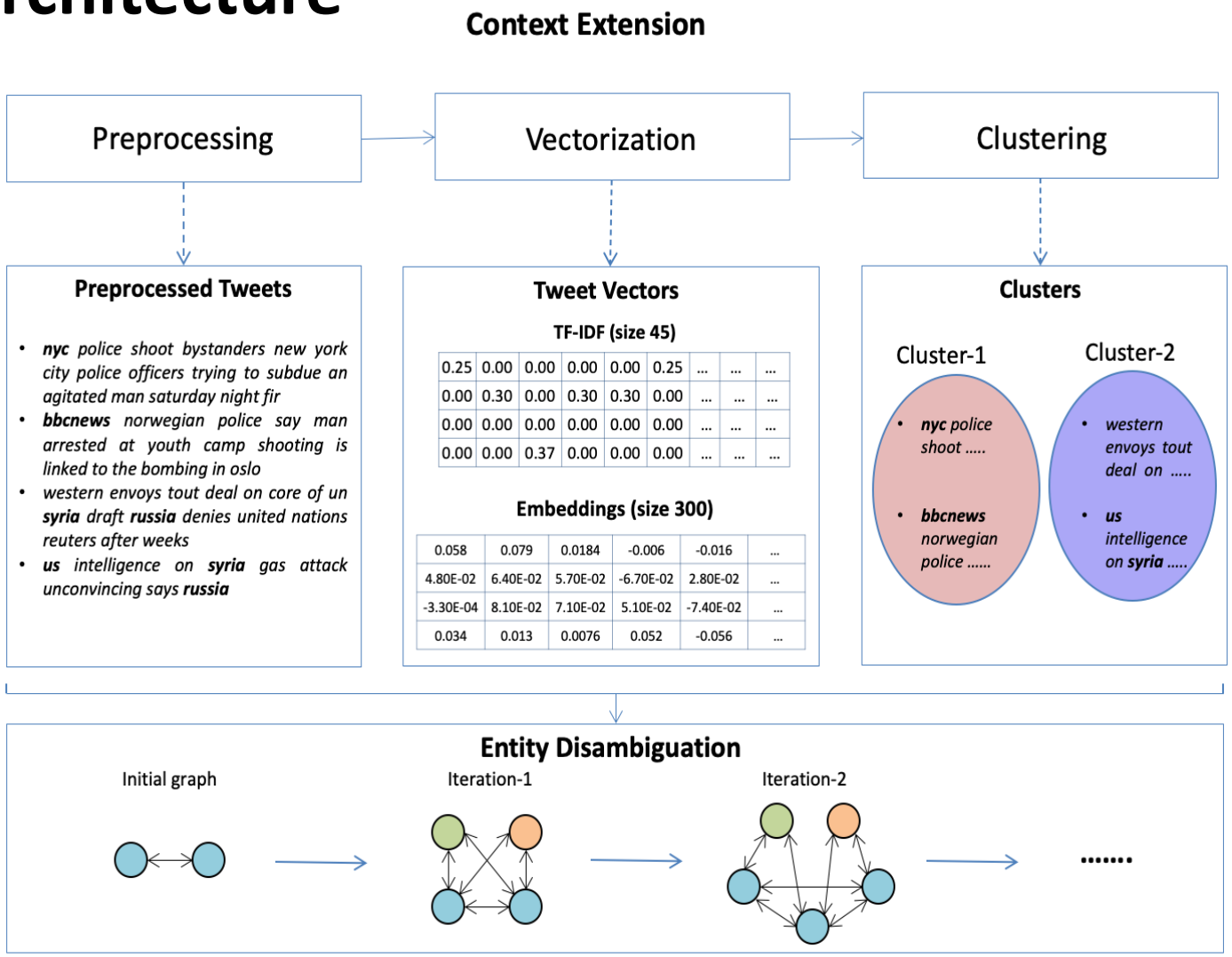
- Extend the context in tweets using unsupervised approaches
- **Preprocessing:** remove emoticons, web URLs, hashtags, punctuations, HTML references and tokenize the tweets
- **Vectorization:** convert tweet text into **TF-IDF** or **Embedding** vectors
- **Clustering:** cluster contextually similar tweets together through unsupervised clustering technique

Entity Disambiguation

- Iterative graph based Entity Disambiguation technique that uses embeddings
- Train a Word2vec model with entities from Wikidata
- Initial graph with entities of unambiguous mentions
- Disambiguate one ambiguous mention in each iteration



GraphEDM Architecture



Experiments & Results

- Details of configurations used in experiments

  - Context Extension Phase
    - Vectorization method : Embeddings
    - Clustering technique : K-Means

  - Entity Disambiguation Phase
    - Ranking algorithm : PageRank
    - Edit distance : Levenshtein distance

Datasets

- Seven gold standard entity disambiguation datasets
- Datasets contain tweet id, list of mentions and mapped entities from the reference Knowledge Base
- **Twitter API** is used to get the tweet text provided the tweet id

Baselines

- Three baselines are used for comparison on all the datasets
- **AIDA:** an online entity disambiguation tool for text and tables
- **WAT-API:** web service for entity disambiguation of text
- **ELTDS:** disambiguation algorithm is performed on dominant entity candidates

Research Questions

- **Q1:** Performance of proposed GraphEDM compared against existing state-of-the-art baseline methods
- **Q2:** Effect of the various clustering techniques
- **Q3:** Effect of the various vectorization techniques

RQ1 : Comparison with Baselines

- GraphEDM outperforms all baselines on five out of seven datasets

- On the other two datasets, we observe competitive results

RQ1 : Comparison with Baselines

Dataset	AIDA			WAT-API			ELTDS			GraphEDM		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Micropost 2014 Test	-	-	41.2	80.6	31.8	45.6	48.4	32.1	38.6	53.3	51.8	**52.5**
Micropost 2014 Train	-	-	50.3	83.3	39.2	53.3	54	30.4	38.9	56.5	55.3	**55.9**
Micropost 2016 Train	-	-	48.5	81.8	47.5	**60.1**	47.9	33.4	39.6	57.4	56.6	57

- High precision is observed in-case of WAT-API
- Mentions in the text differ from surface forms in the Knowledge Base

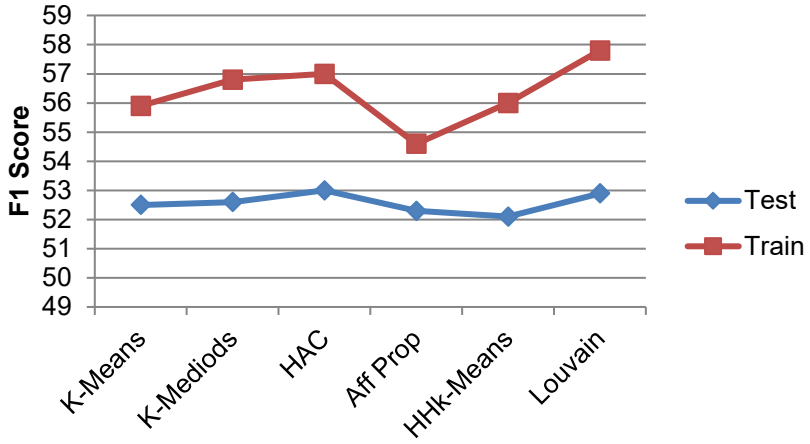
RQ2 : Effect of Clustering

Unsupervised clustering techniques

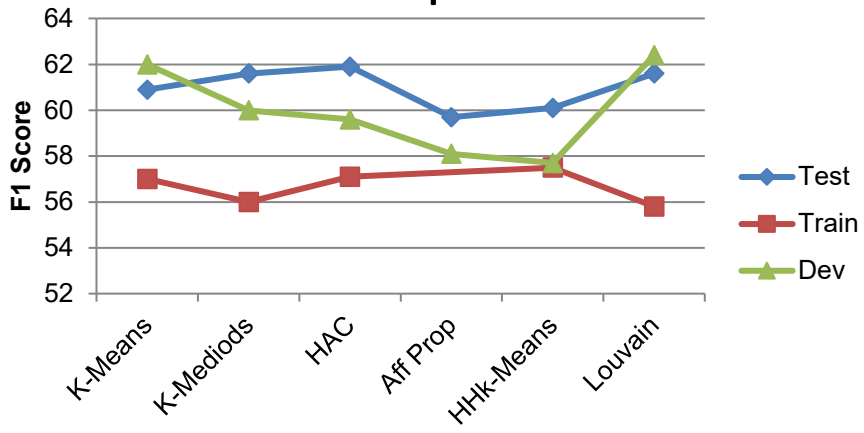
- K-Means
- K-Medoids
- Hierarchical Agglomerative
- Affinity Propagation
- Hybrid Hierarchical K-Means
- Louvain

RQ2 : Effect of Clustering

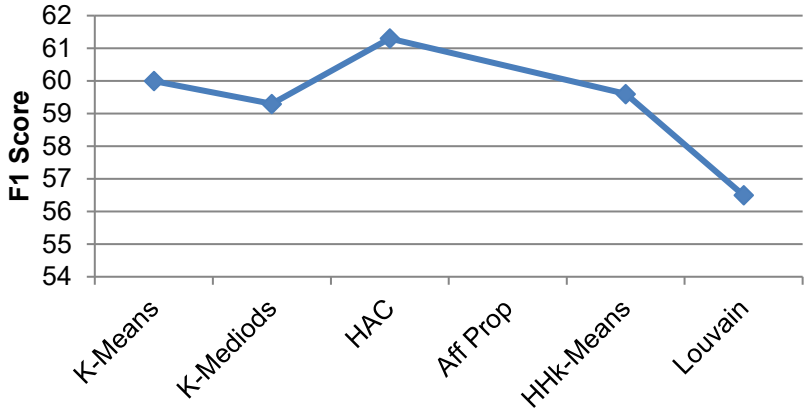
Micropost 2014



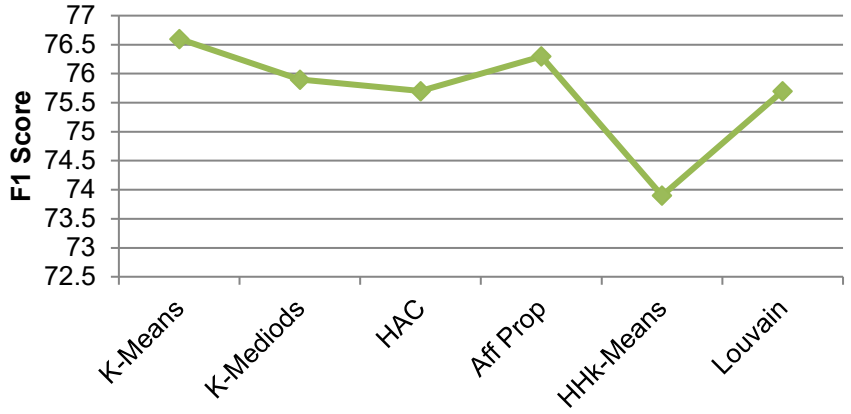
Micropost 2016



Brian Collection

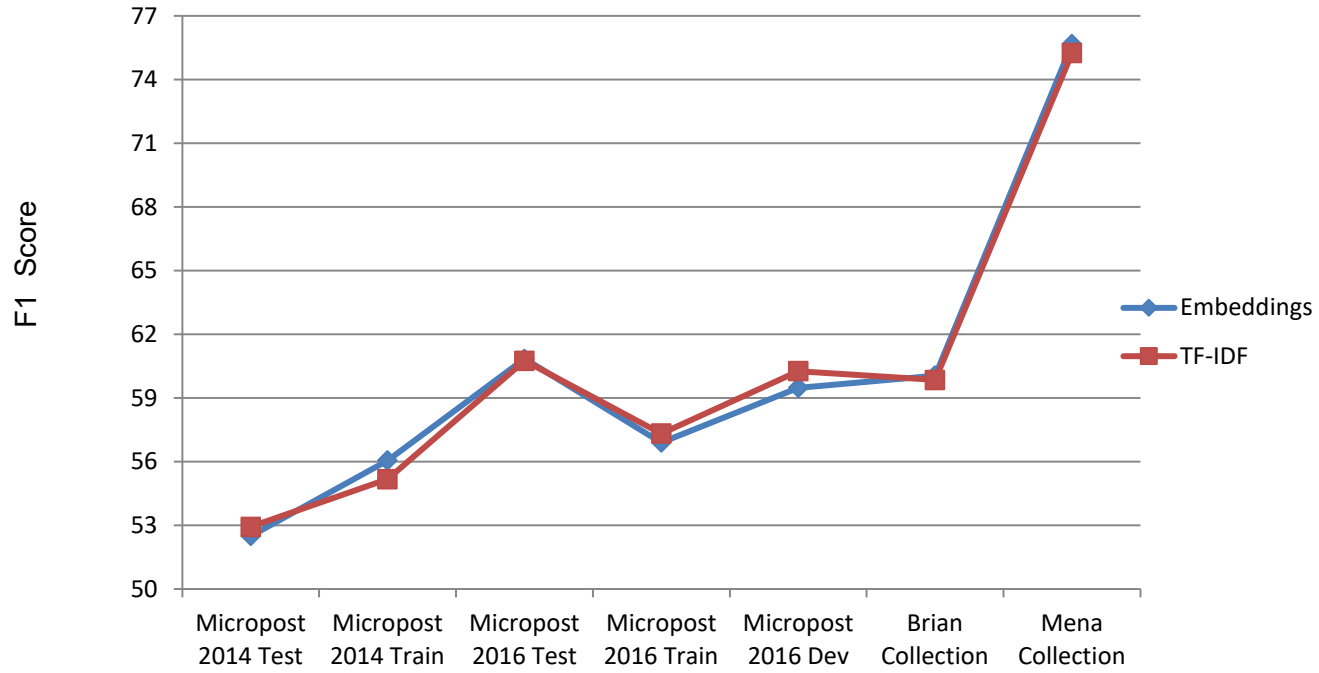


Mena Collection



RQ3 : Effect of Vectorization

- Vectorization techniques
  - TF-IDF: importance of a word in a document in a corpus
  - Embeddings : based on distributional hypothesis



Conclusion & Future work

- We propose GraphEDM, an effective framework for entity disambiguation in microposts.
- We show that the context enhancement in tweet datasets can improve the performance of disambiguation
- We observe that our framework for entity disambiguation outperforms the SoTA baselines by up to 15.13%
- In future, we would like to study the effect of dataset's size and dynamic clustering

SDS2021

SWISS CONFERENCE
ON DATA SCIENCE

UNI
FR

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Thank You!