

# GraphEDM: A Graph Based Approach for Entity Disambiguation in Microposts

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

## Table of Contents

- Introduction
- Related work
- Methodology
- Experiments & Results
- Conclusion & Future work

## 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

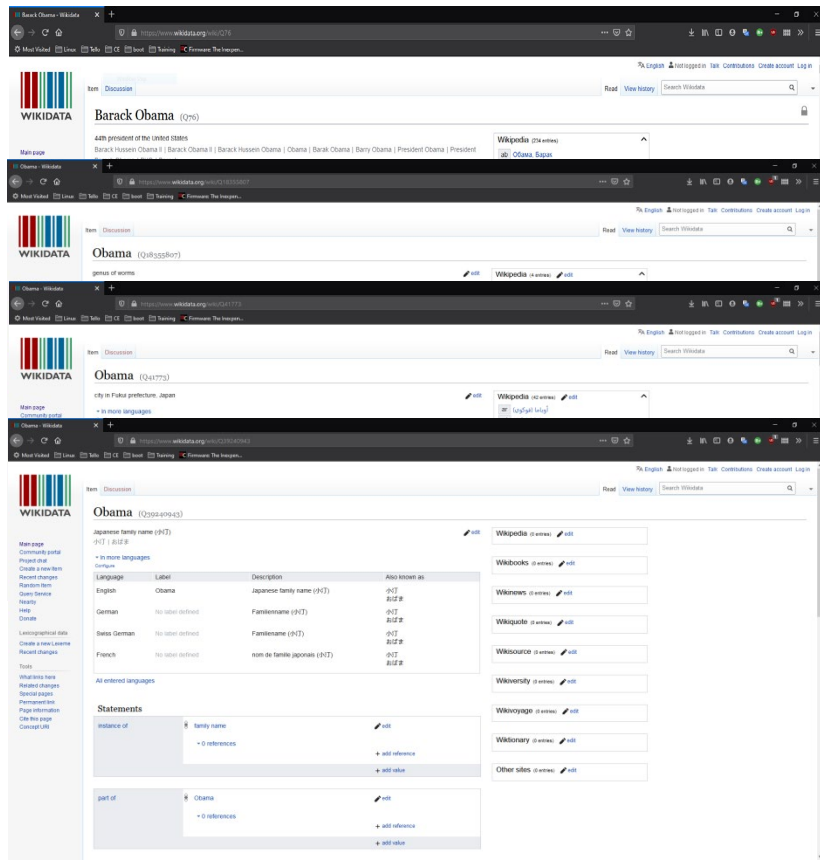
The screenshot shows the Wikidata page for Barack Obama. The main content area displays a table with columns for Language, Label, Description, and Also known as. The table lists various language labels for Barack Obama, including English, German, Swiss German, and French. The English label is 'Barack Obama' and the description is '44th president of the United States'. The 'Also known as' column lists various names and titles, such as 'Barack Hussein Obama II', 'Barack H. Obama', 'President Obama', and 'US President Barack Obama'.

Language	Label	Description	Also known as
English	Barack Obama	44th president of the United States	Barack Hussein Obama II Barack H. Obama Barack Hussein Obama Obama Barack Obama Barry Obama President Obama President Barack Obama Barack Obama Barack Obama
German	Barack Obama	44. Präsident der Vereinigten Staaten	Barack Hussein Obama, II Obama Barack H. Obama Barack H. Obama Präsident Obama Präsident Barack Obama US-Präsident Barack Obama
Swiss German	Barack Obama	No description defined	
French	Barack Obama	président des États-Unis de 2008 à 2017	Barack Hussein Obama Barack Hussein Obama II Obama

The screenshot shows the Wikidata page for New York City. The main content area displays a table with columns for Language, Label, Description, and Also known as. The table lists various language labels for New York City, including English, German, Swiss German, and French. The English label is 'New York City' and the description is 'largest city in the United States'. The 'Also known as' column lists various names and titles, such as 'NYC', 'The Big Apple', 'The City of New York', and 'The Empire State'.

Language	Label	Description	Also known as
English	New York City	largest city in the United States	NYC New York The Big Apple Big Apple City of New York NY City New York, New York New York City, New York New York, NY
German	New York City	Metropole an der Ostküste der Vereinigten Staaten	Big Apple NYC New York, Stadt New York New-York-Stadt
Swiss German	New York City	No description defined	
French	New York	ville la plus peuplée des États-Unis	New York City la Grande Pomme NYC New York

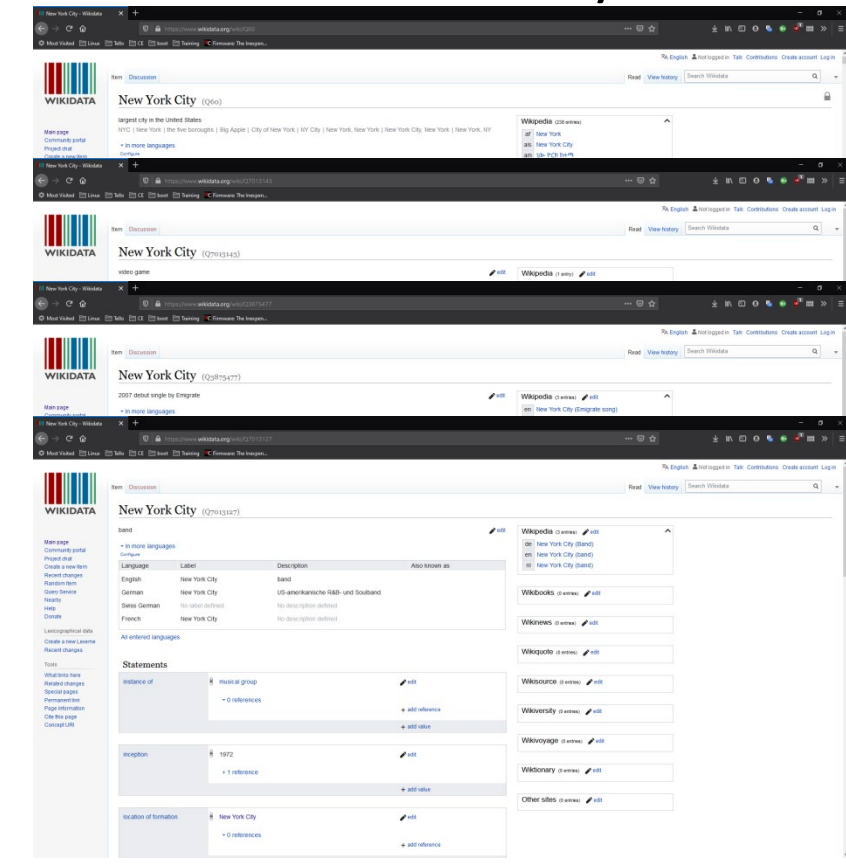
# Need for Disambiguation Obama



....  
....

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

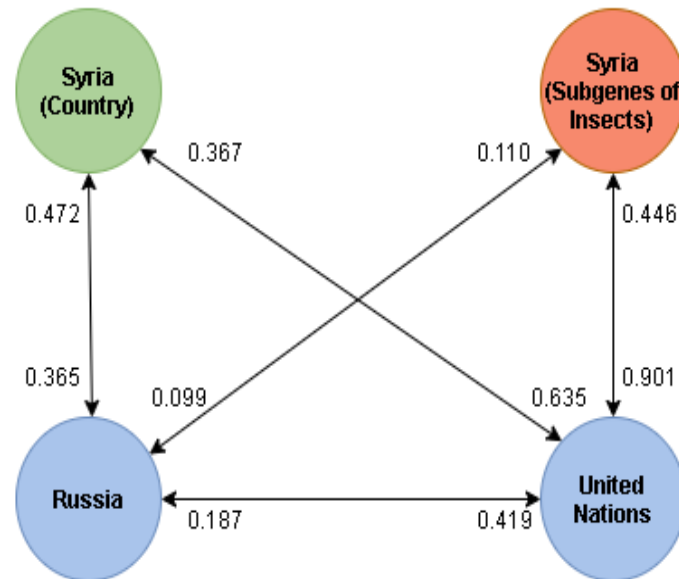


# Context Extension

- 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

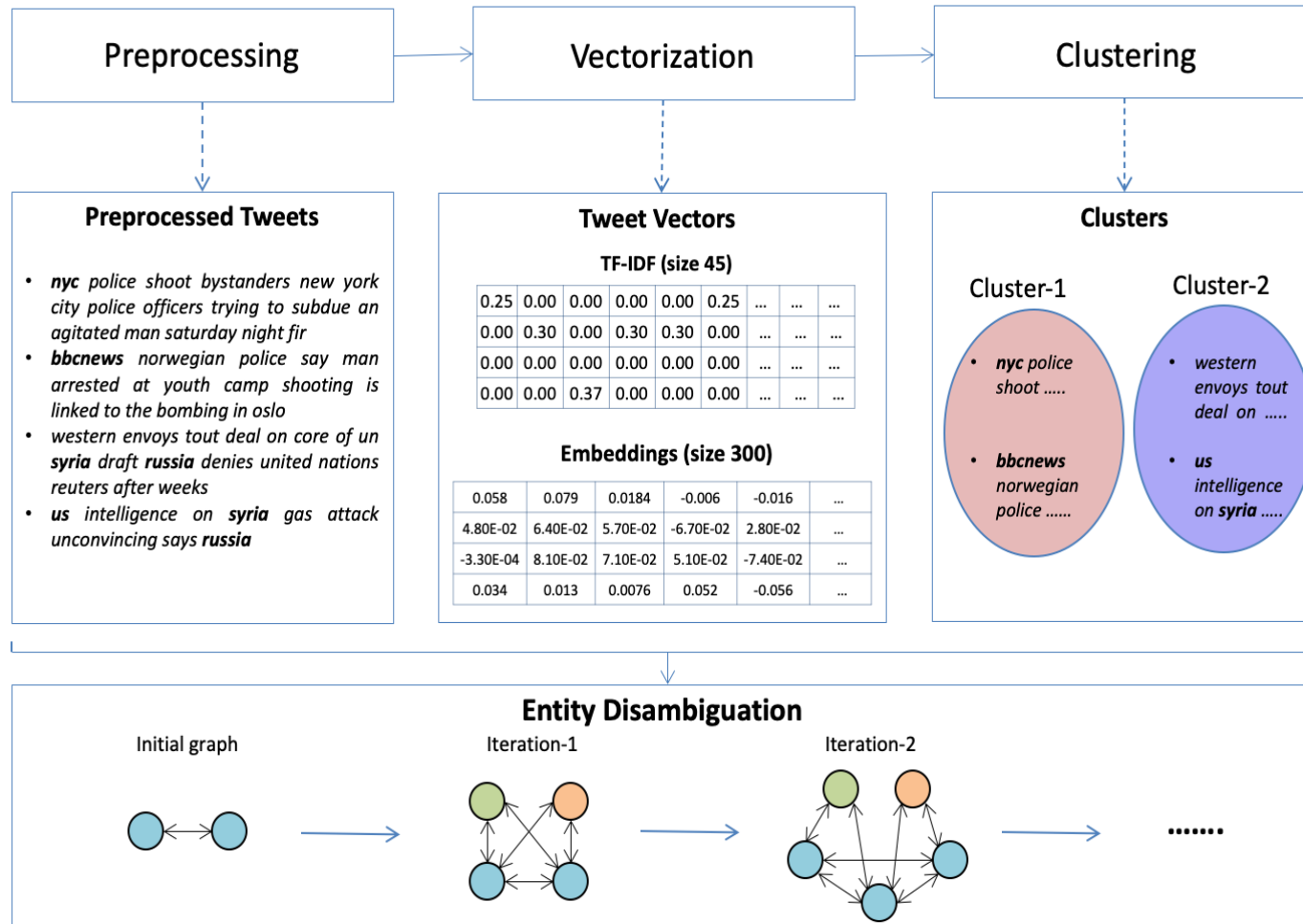
# Entity Disambiguation

- 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

## Context Extension



# 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	<b>52.5</b>
Micropost 2014 Train	-	-	50.3	83.3	39.2	53.3	54	30.4	38.9	56.5	55.3	<b>55.9</b>
Micropost 2016 Train	-	-	48.5	81.8	47.5	<b>60.1</b>	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

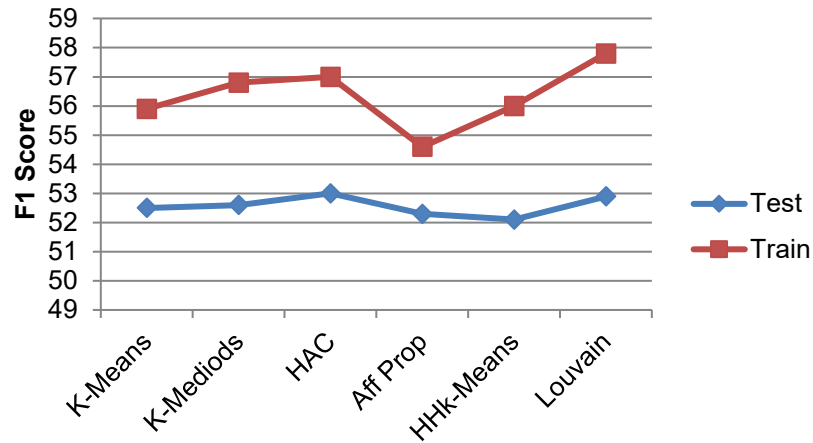
## RQ2 : Effect of Clustering

Unsupervised clustering techniques

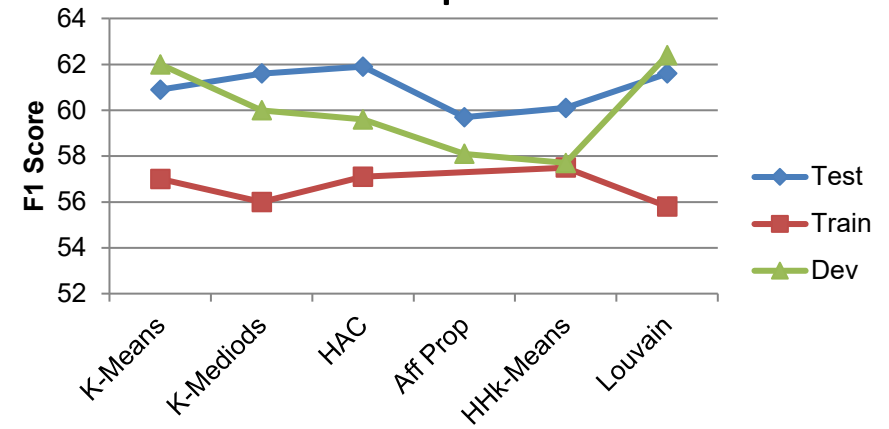
- K-Means
- K-Medoids
- Hierarchical Agglomerative
- Affinity Propagation
- Hybrid Hierarchical K-Means
- Louvain

# RQ2 : Effect of Clustering

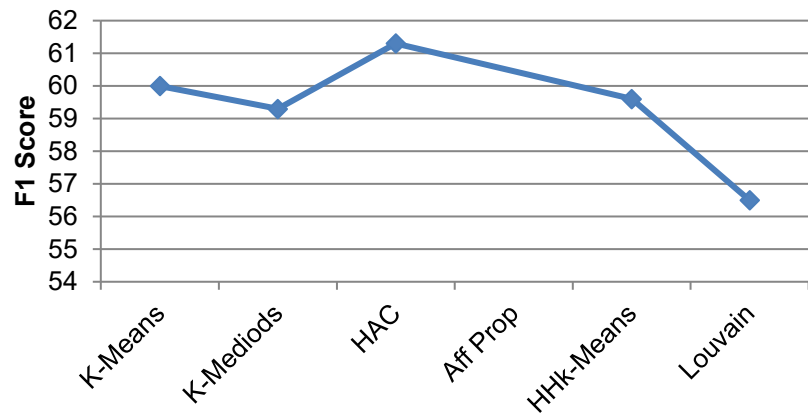
## Micropost 2014



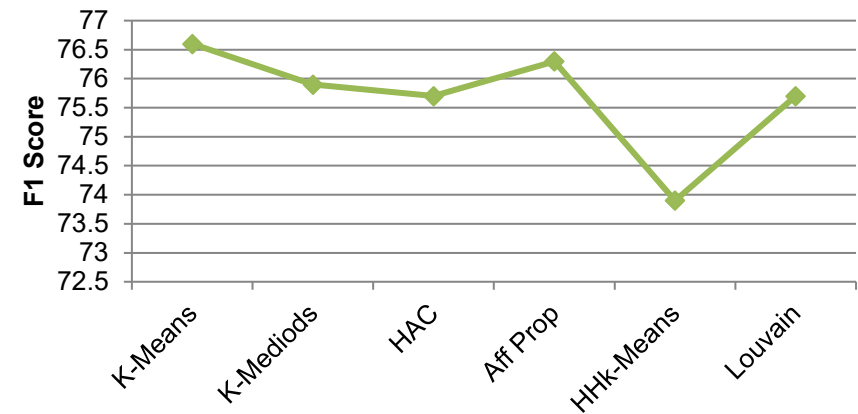
## Micropost 2016



## Brian Collection

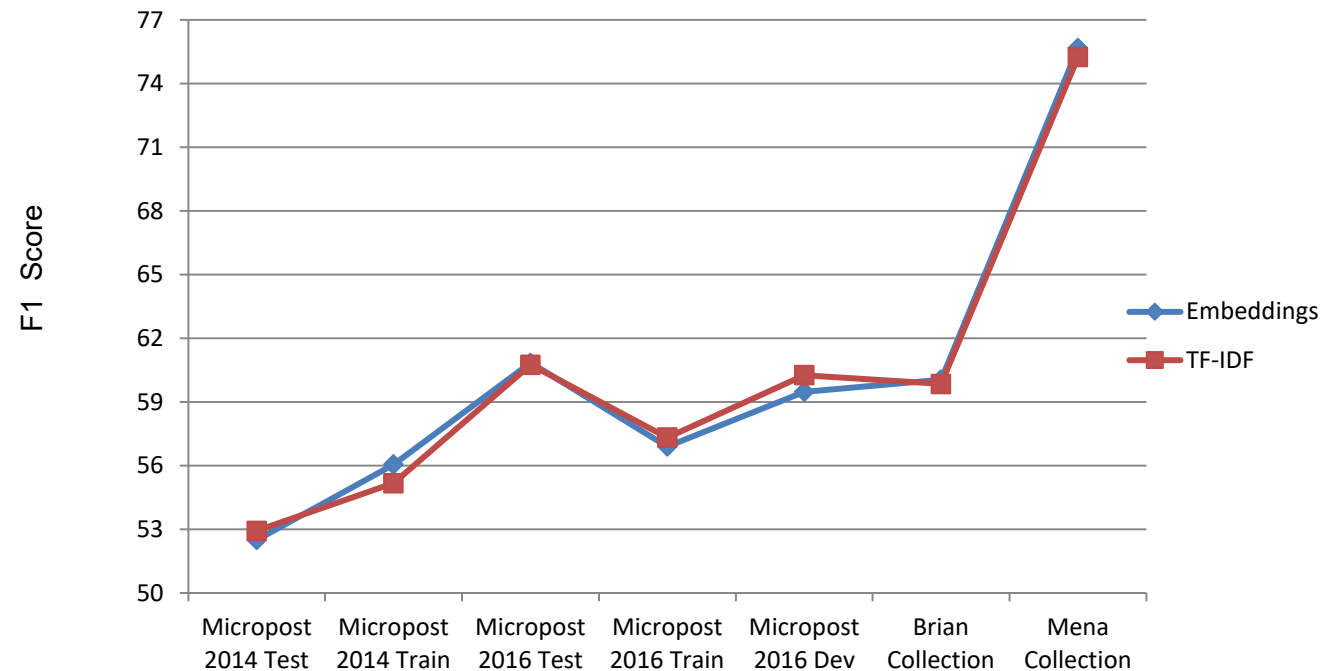


## Mena Collection



## 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

**Thank You!**