

# AIDA-Social: Entity Linking on the Social Stream

Yusra Ibrahim, Mohamed Amir Yosef, Gerhard Weikum  
 {yibrahim|mamir|weikum}@mpi-inf.mpg.de

## Motivation

- The tremendous increase in social media popularity drove an abundance of user-generated content.
- Adding semantics to this content assists in subsequent Information Retrieval tasks such as relation extraction and semantic search.
- Named Entity Linking (NEL) disambiguates names to their corresponding canonical entities in Knowledge Bases such as YAGO.
- NEL in social media, more specifically in microblogs, is a challenging task due to the brevity, lack of contextual information, and time-varying importance of entities.
- Twitter is the most popular microblog with 500 million Tweets per day.

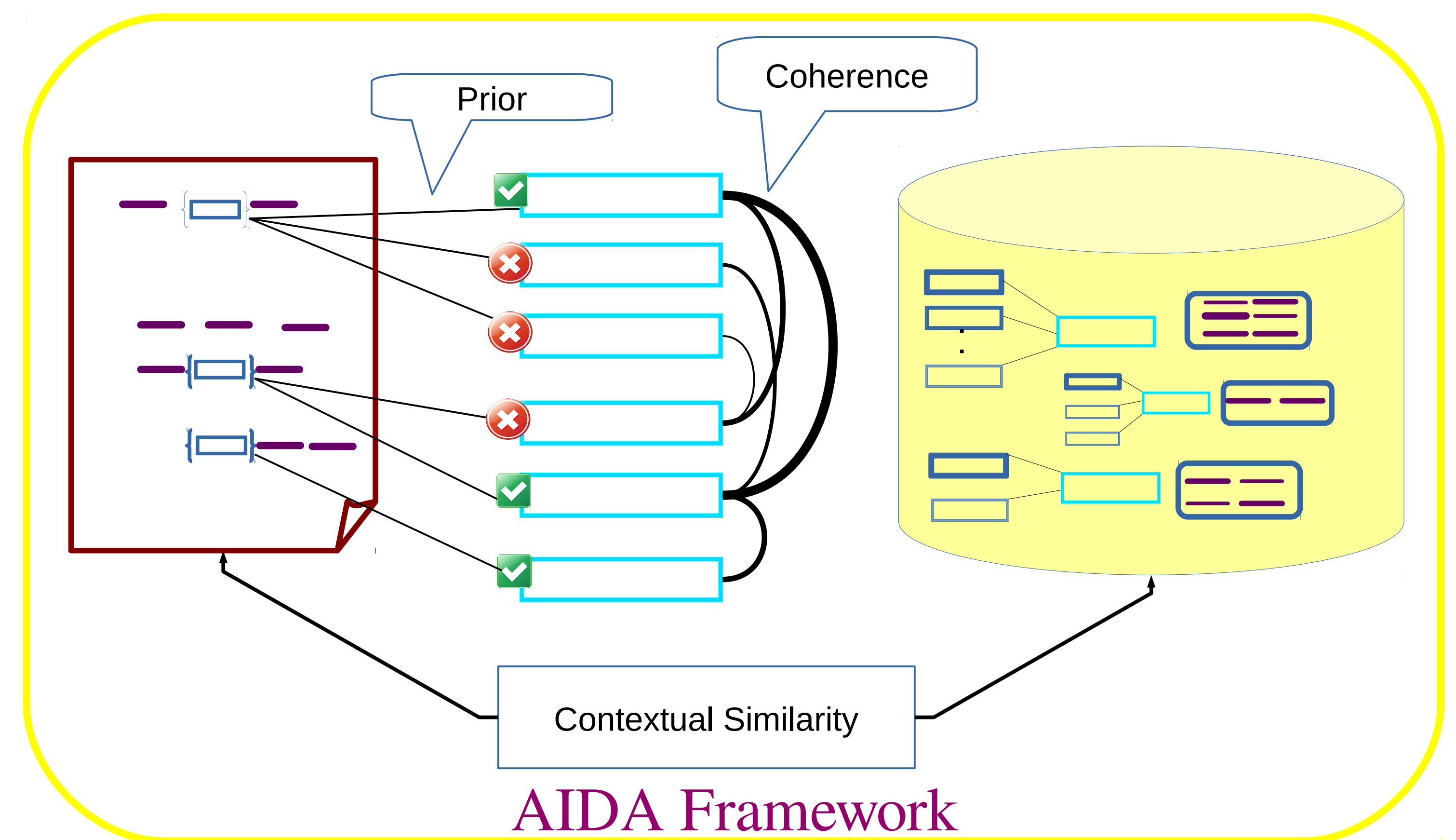
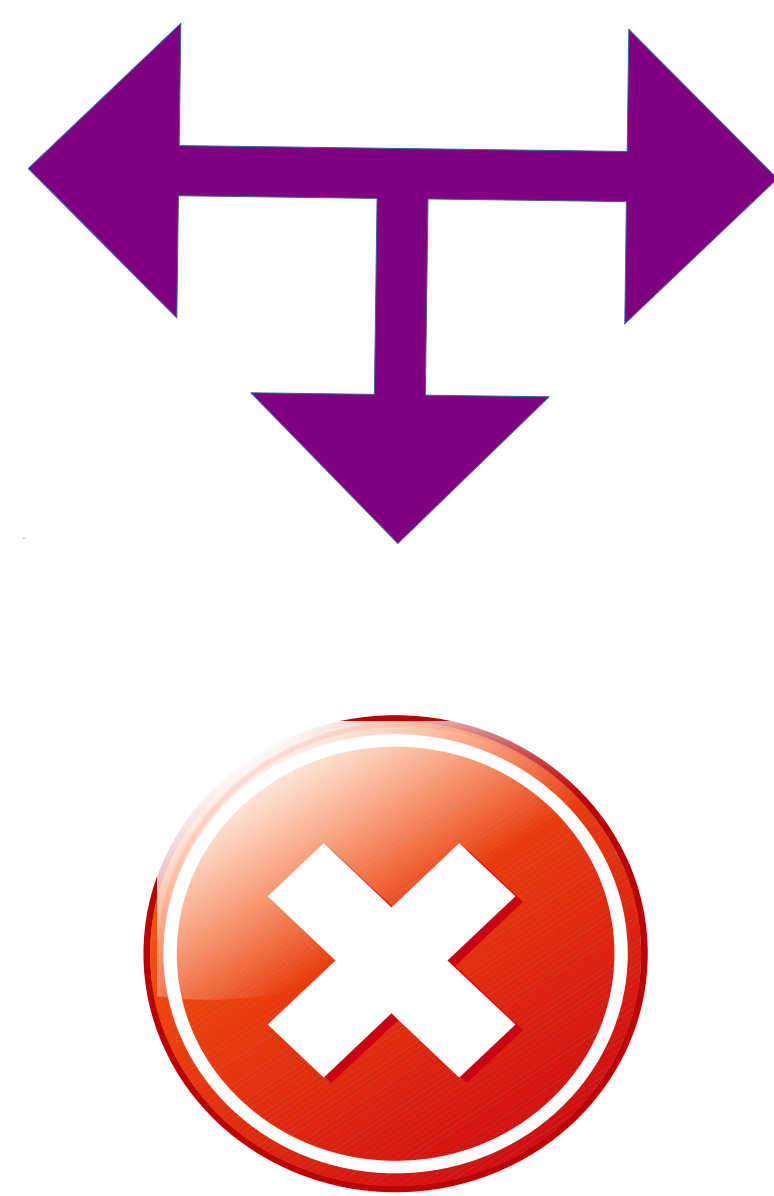
## Challenges

**Sample Tweets**

*Short, no context*  
 [[Mueller]] Goes South of Border For Extreme Rehab  
<http://bit.ly/rfd3GN>

*Cryptic Mentions*  
 I think [[@joey7barto]] was in the wrong today!  
 He should not have got involved  
 and assaulted Gerviniho like that #justsaying

*Time-Varying Entity Importance*  
 Such an unbelievable loss. RIP [[Amy]]..xxx we love you.. Xx



## Proposed Techniques

**Mention Normalization**

Build a normalized set of mentions for every Cryptic mention

Examples of normalized mentions:

- #tagdef
- Joseph Barton
- I think @joey7barto was in the wrong today! He should not have got involved and assaulted Gerviniho like that #justsaying
- @JLSOfficial #JLSquestiontime do you know a country like Lithuania? and do you know that u have fans in there. One of them is me-Smile.:) x
- I'm fight cancer now, so I can fight for families in September: <http://ndp.ca/hhJyz> #NDP #Cdnpol.

**Temporal Importance**

Wikipedia article traffic statistics

Amy Winehouse has been viewed 9811788 times in 201107. This article ranked 2096 in traffic on en.wikipedia.org.

23 July 2011

Such an unbelievable loss. RIP Amy..xxx we love you.. Xx

Amy Wright vs. Amy Winehouse  
 116 vs. 4231460

Calculate Entity importance based on the micropost's publication date and time.

**Contextual Enrichment**

Add Extra Context to the micropost

#tagdef + RT @DMVFollowers #RetweetThisIf you've been living in the #DMV for more than 2 years!

User Profile

URLS

Mueller Goes South of Border For Extreme Rehab  
<http://bit.ly/rfd3GN>

MUELLER Headed South of Border For Extreme Rehab

Brooke Mueller is going to Mexico to ingest a bunch of hallucinogens -- but not for fun -- to break her addictions.

Chiqui Cancun

Clustering Using DBSCAN with three overlapping coefficient: #tags, @userMentions, and normal tokens

Collective NEL + Full Content + Entities + Frequent Words

## Experimental Results

	Mention Normalization	Temporal Importance	Contextual Enrichment	P@1
#Microposts2014 annotated corpus	✗	✗	✗	58.41%
Filter out: numbers, General Concepts, OOKB	✗	✗	✗	
Tweets 2809	✓	✗	✗	61.07%
Mentions 3392	✗	✓	✗	67.35%
Unique Mentions 2019	✗	✗	✓	62.98%
@ 506	✓	✗	✓	69.69%
# 1085	✓	✓	✗	66.00%
Entities 1683	✓	✓	✓	69.07%
Baseline: AIDA - Prior + Contextual Similarity	✗	✓	✓	72.19%

## Conclusion and Future Work

- ✓ Temporal Entity Importance, Contextual Enrichment, and Mention Normalization yields +13% gain over P@1.
- ✓ Adding extra context assists in mention-entity similarity measures
- ✓ Temporal Importance improves the accuracy of NEL, specially on plain tweets.
- ✓ Adding context from URLS yields better gain than adding context from #Tags and @userMentions
- ✓ Clustering assists in coherence measures.
- TODO: Other Datasets
- TODO: Other ways to estimate Temporal Importance
- TODO: More on Clustering

