

# Tracking the Diffusion of Named Entities

TBD

## Abstract

To do

## Index Terms

To do

## I. INTRODUCTION

The aim of this paper is to understand how named entities *emerge* and *spread* through social media based discourse. We are interested in exploring the following research questions:

- 1) **RQ1:** How can we accurately detect named entities in social media based discourse, given its myriad formats, often informal vernacular, and inherent noise (e.g. misspellings, abbreviations, etc.)?
- 2) **RQ2:** Under what conditions do entity mentions diffuse through discourse? And when are people *most likely* to be influenced into then discussing entities?
- 3) **RQ3:** How can we predict the discussion of certain named entities and who will begin talking about them?

## II. DATASETS

For this research we will use the following datasets:

- 1) Reddit data – download and access all of the data from the full dump.<sup>1</sup>
- 2) CoNLL 2003 data – a corpus of newswire texts, annotated for named entity chunks and types. This describes where entity mentions are in the text, including locations, organisations, and person mentions.
- 3) Twitter data; unannotated – we have a large corpus of English tweets that we can use here.
- 4) Twitter data; annotated – there are two datasets annotated with named entities. These are from Ritter’s 2011 EMNLP paper, and the W-NUT 2015 shared task.

## III. RESEARCH STAGES

### A. Stage 0: Data Preparation and NER

-To do:

- Annotate corpora with detected entities using basic typing of: person, location, organisation
- Run NER software over dataset and validate accuracy of this (using basic measures)
- Run NER over entire dataset to extract entities

<sup>1</sup>[https://archive.org/details/2015\\_reddit\\_comments\\_corpus](https://archive.org/details/2015_reddit_comments_corpus)

### B. Stage 1: Exploratory Analysis

-To do:

- Plot relative frequency distribution as a function of time for named entities, and characterise the *shape* of the entities
- Apply lifecycle model to profile users' NER citations over time and investigate how users' profiles are influenced by global, community, and prior behaviour dynamics

### C. Stage 2: Diffusion Analysis

-To do:

- Model the spread of named entities through user profiles (could use multivariate diffusion models here)

### D. Stage 3: Forecasting

-To do:

- Implement models to forecast if a user will mention an entity and who that will be (hard!)


## IV. DATA PREPARATION AND NER

To conduct our study, we need to convert 140GB of compressed Reddit posts into a set of interlinked and time-ordered conversations and the entities mentioned in each of them. This provides a number of sub-challenges: sampling of the Reddit data, creating a linked series of conversations, and picking out entity mentions in this text type. Reddit data is largely unexplored in the NLP community, despite the large volume of it and the especially rich metadata. This poses additional challenges: certainly, given the lack of work on Reddit text, there are no annotated datasets available yet, so supervised in-domain work is not directly possible yet. Additionally, the datasets are large, which makes it important to choose a good subset of data on which to do prototyping and development, in order to keep research cycles short. The result that we come to at the end of this stage is a rich dataset for tracking entity and concept diffusion within and across communities.

The Reddit dataset [1] is comprised of a sequence of comments, with one JSON record for each one. These are ordered temporally. Reddit itself is roughly similar to a forum, where top-level divisions are made by topic. Within each topic, or *subreddit*, there are posts, which begin with either a short piece of text or a link to an external resource – typically an image, video, or interesting article. Users then may publish comments for each post, and reply to each others' comments. This leads to a threaded discussion, centred on a particular topic, with a hierarchical comment structure (see Figure IV).

NER:

- blended text type approach as in ritter 2011
- CRF w/ classical features and all brown string depth features
- use entity-recognition toolkit
- describe reddit annotation/tuning annotation approach (skim, find things it got wrong, annotate them, add to set)




# /R/MILDLYINTERESTING

FEED YOUR MILD SIDE

[comments](#) [related](#)

↑  
1760  
↓



**Beirut gas station with hanging garden** (imgur.cc)  
submitted 3 hours ago by [catloveroftheweek](#)  
45 comments [share](#)

all 45 comments

sorted by: [best](#) ▼

↑ ↓ [-] [SuaveAndAloof](#) 29 points 53 minutes ago

That gas station is probably the most aesthetically pleasing one in says in the background) a pretty popular bar hopping area. The w and it basically just loops around the metal wires.

[permalink](#)

↑ ↓ [-] [Superflypirate](#) 57 points an hour ago

For some reason that doesn't seem safe.

[permalink](#)

↑ ↓ [-] [frekinghell](#) 24 points 41 minutes ago

For some reason it seems like it's very pleasing *and* safe.

[permalink](#) [parent](#)

↑ ↓ [-] [moeburn](#) 2 points 12 minutes ago

Lotsa bugs though.

[permalink](#) [parent](#)

↑ ↓ [-] [BobNelsonUSA1939](#) 1 point 4 minutes ago

Until the bombs start going off.

[permalink](#) [parent](#)

↑ ↓ [-] [Nothingbutfreewill](#) 7 points 41 minutes ago

I got the same feeling too, as if something is going to catch on

[permalink](#) [parent](#)

↑ ↓ [-] [PM\\_ME\\_UR\\_BEST\\_TRAIT](#) 10 points 27 minutes ago

Well, no matter what the gas station looks like, you do not there is something there to spark the flame in the first plac

[permalink](#) [parent](#)

- estimate Brown c, w based on impact paper findings
- extract reddit tuning set
- tune twitter / newswire balance
- one, three, four, ten entity classes?
- prefer precision? which beta value?

#### REFERENCES

- [1] J. Baumgartner, "Complete public reddit comments corpus," 2015.