# Prototype Plan

## Prototype Deliverable

A solution that can demographically group the sentiment analysis results of users on Twitter

Demographic groups:

* Male / Female
* Age bracket (18-30 etc)
* City (1 of 3, else ignore)

1. Birmingham
2. London
3. Liverpool

The solution therefore makes it easy to determine that, for example, Males aged 18-30 from London show the following emotions towards #generalelection

A large number of uncategorised results would suggest strong Twitter API protection and reduced ability to infer demographic data.

## Demographic Inference

How to determine **gender**?

* Name -> gender matching

How to determine **age?**

* Bio search for age
* Tweet mentions with happy nth birthday
* ~~Born on from profile~~

How to determine **location**?

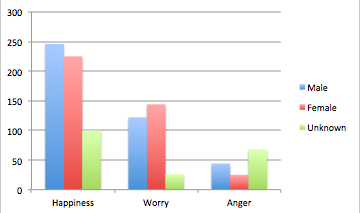
* Twitter profile location explicit and/or bio
  + Not an inference method, but one used as a simplistic way to demonstrate inference of age and gender
* You could mention the potential for location inference using the various methods discussed in “A Survey of Location Prediction on Twitter”

## Evaluation Key Statistics

Step-by-step walkthrough:

1. For a given hashtag, such as #generalelection, retrieve 1000 recent tweets
2. Sentiment analyse every tweet to produce a relative % emotion displayed
3. For every one of those tweets, perform demographic analysis, as above
4. Categorise sentiment according to each of the demographics SEPARATELY, with anybody who couldn’t be categorised being ‘unknown’
5. The % of tweeters from the 1000 sample that end up in the ‘unknown’ category for each demographic will provide you with the basis for making your evaluation as to the extent by which privacy is secure

For example, the **gender** graph from the 1000 sample may appear as follows:



|  |  |  |  |
| --- | --- | --- | --- |
| Gender | Happiness | Worry | Anger |
| Male | 246 | 122 | 44 |
| Female | 225 | 144 | 25 |
| Unknown | 100 | 26 | 68 |

The table…

The number 246 in X = happiness and Y = Male means

246 out of the 1000 sample displayed happiness as the predominant emotion towards the given topic.

Hence the number of people with ‘unknown’ gender is the margin for privacy being protected… although for gender this is likely to just be because their twitter name is gibberish.