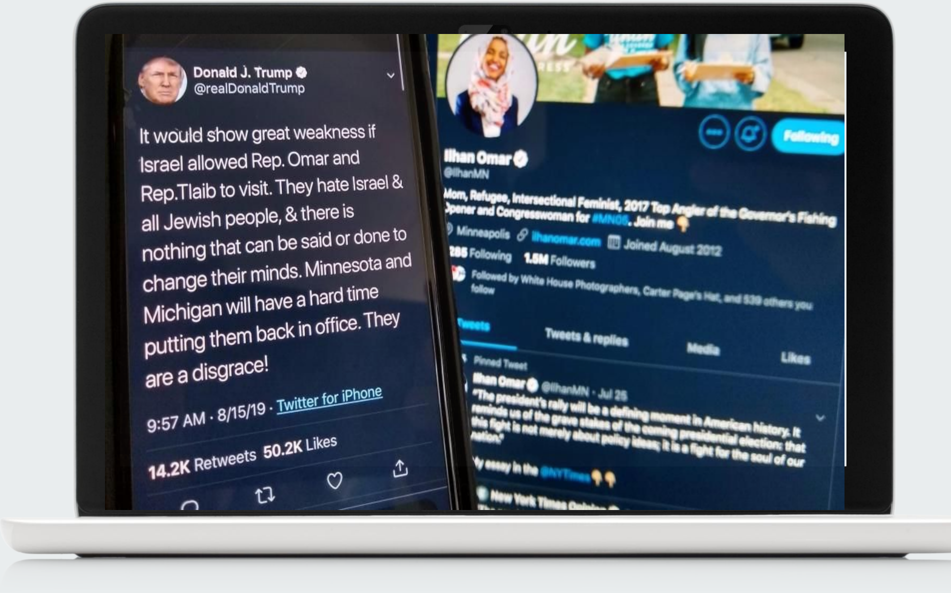
# ``Share Love not Hate'' - Assessing Hate Speech Detection methods using TF-IDF and POS tagging approach on Twitter

Kshitija Hande
(1131778)



Donald J. Trump
@realDonaldTrump

It would show great weakness if Israel allowed Rep. Omar and Rep.Tlaib to visit. They hate Israel & all Jewish people, & there is nothing that can be said or done to change their minds. Minnesota and Michigan will have a hard time putting them back in office. They are a disgrace!

9:57 AM · 8/15/19 · Twitter for iPhone

14.2K Retweets 50.2K Likes

Ilhan Omar
@IlhanMN

Mom, Refugee, Intersectional Feminist, 2017 Top Angler of the Governor's Fishing Opener and Congresswoman for #MN05. Join me

Minneapolis ilhanomar.com Joined August 2012

285 Following 1.5M Followers

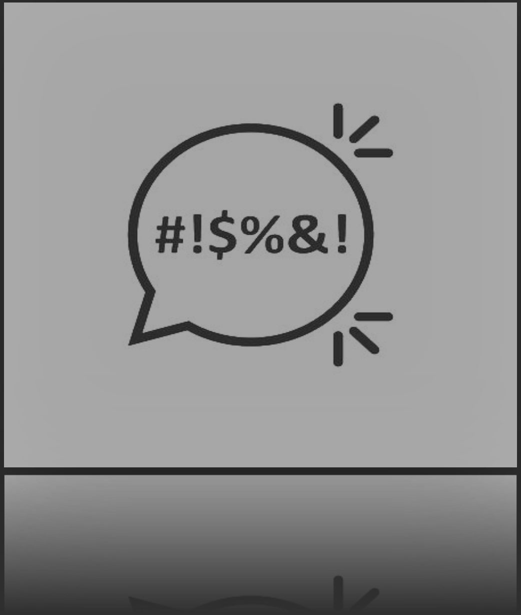
Followed by White House Photographers, Carter Page's Hat, and 539 others you follow

Ilhan Omar @IlhanMN · Jul 25
"The president's rally will be a defining moment in American history. It reminds us of the grave stakes of the coming presidential election: that this fight is not merely about policy ideas; it is a fight for the soul of our nation."

My essay in the @NYTimes

New York Times Opinion

# Outline



The Problem

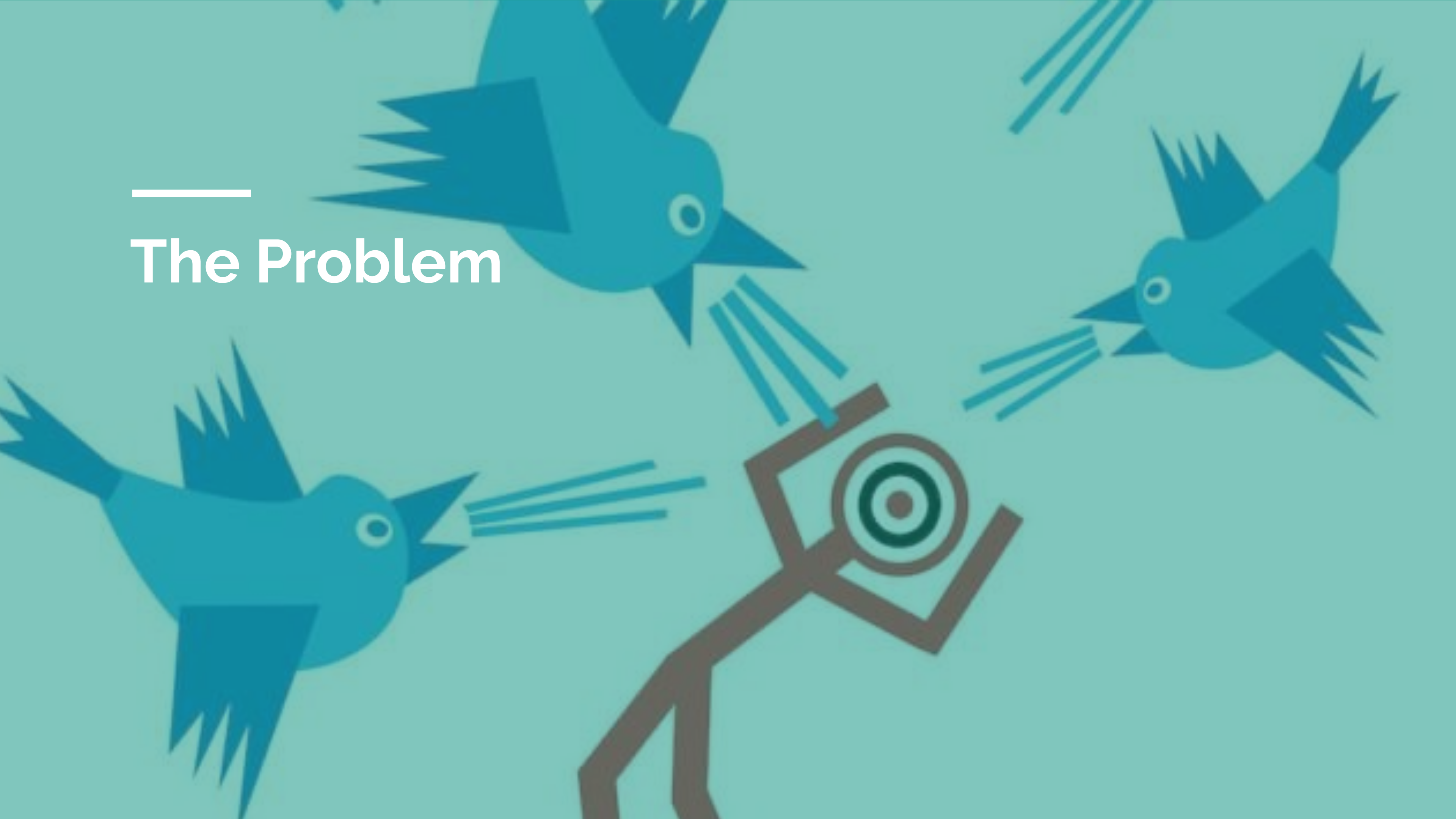
Methodology

Results

Limitation

Conclusion

# The Problem

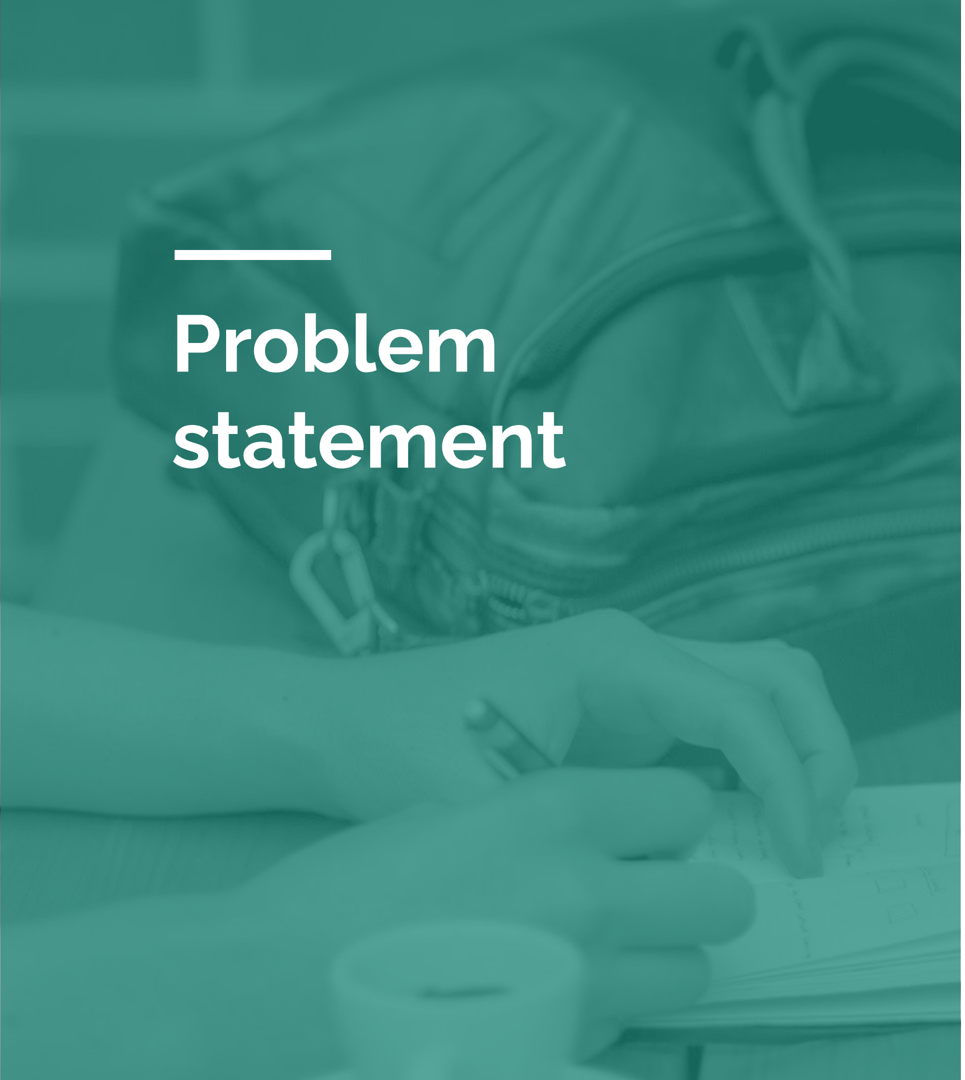


# Why do we care?

Exponential increase in the use of the internet by people of different cultures, sexual orientation, ethnicities and educational backgrounds.

Differentiating hate speech and offensive language is a key challenge.

New York Times reported personal attacks motivated by bias and prejudice



# Problem statement

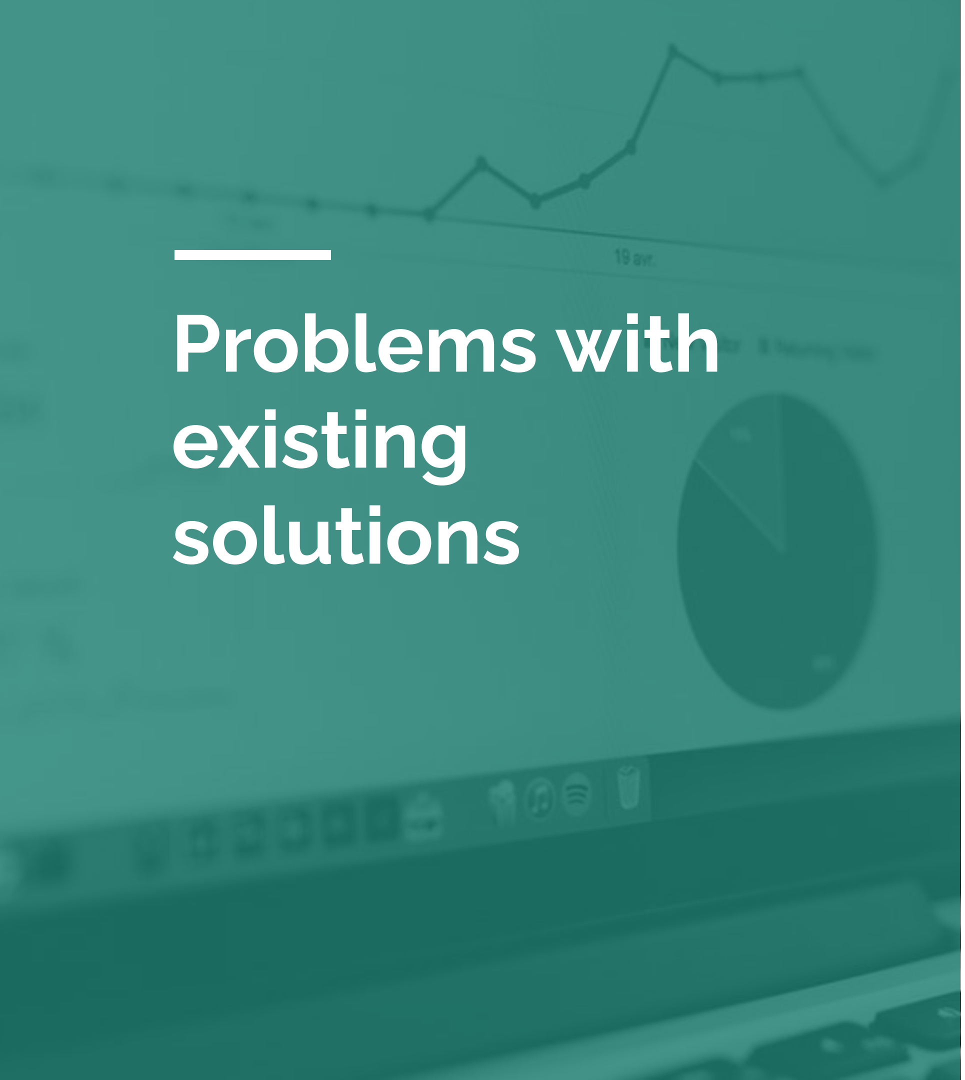
Classification of tweets on Twitter into three classes: hate speech, offensive language and neither.
We perform experiments by leveraging TF-IDF and POS tagging features as input to machine learning models.

# Challenges

The style of social media especially hateful data is different

Uses abbreviations and spelling variations

Words cannot be taken literally.



# Problems with existing solutions

OOV words are impossible to train

Example : #BanPUBG

Issue with pre-trained embeddings

Splitting tweets to normalize input

Training on domain-specific data

# Methodology

# Data

01

## DATASET DISTRIBUTION

| DATASET | COUNT |
|---|---|
| Train | 14,869 |
| Validation | 3,718 |
| Test | 6,196 |
| TOTAL | 24,783 |

# Data

02

## VALUES IN THE CLASS COLUMN

0:"hate speech"

1:"offensive language"

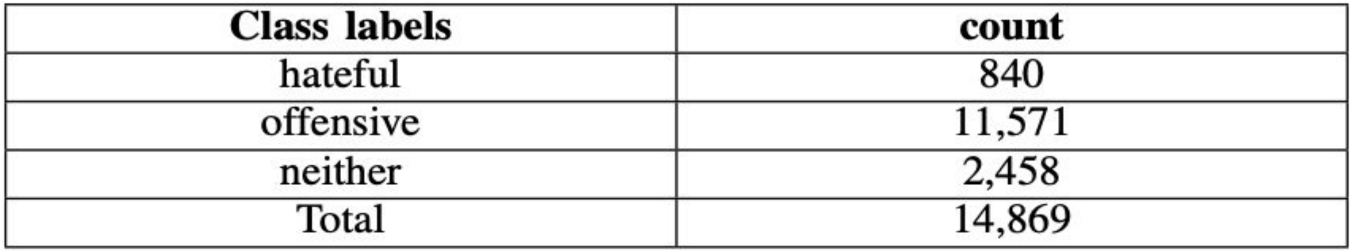
2:"neither"

## COLUMNS IN THE DATASET

| |
|---|
| count |
| hate_speech_count |
| offensive_language_count |
| neither_count |
| class |
| tweet |

# Data

03



| Class labels | count |
|---|---|
| hateful | 840 |
| offensive | 11,571 |
| neither | 2,458 |
| Total | 14,869 |

TABLE I

CLASS DISTRIBUTION IN TRAINING SET

# Data

04

- Data Preprocessing :
  - Lowercase
  - Stemming
  - Lemmatization
- Data cleaning
  - Stopwords
  - URLs
  - Twitter Mentions
  - Retweet Symbols
  - Emojis

# Feature Extraction

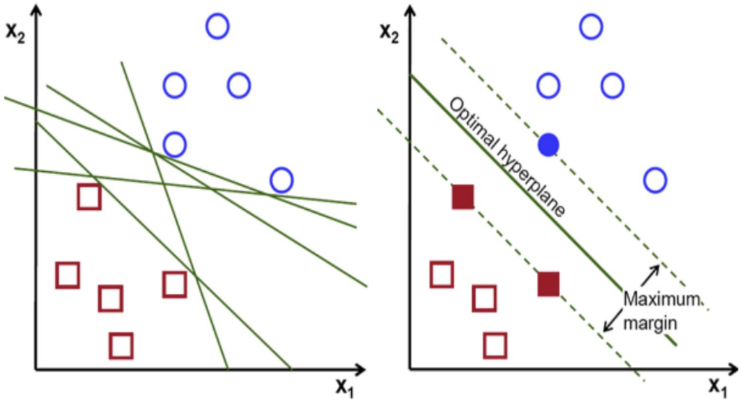
```
tfidf(d, t) =     tf(t) *  idf(d,t)
```

# Feature Selection - Logistic Regression

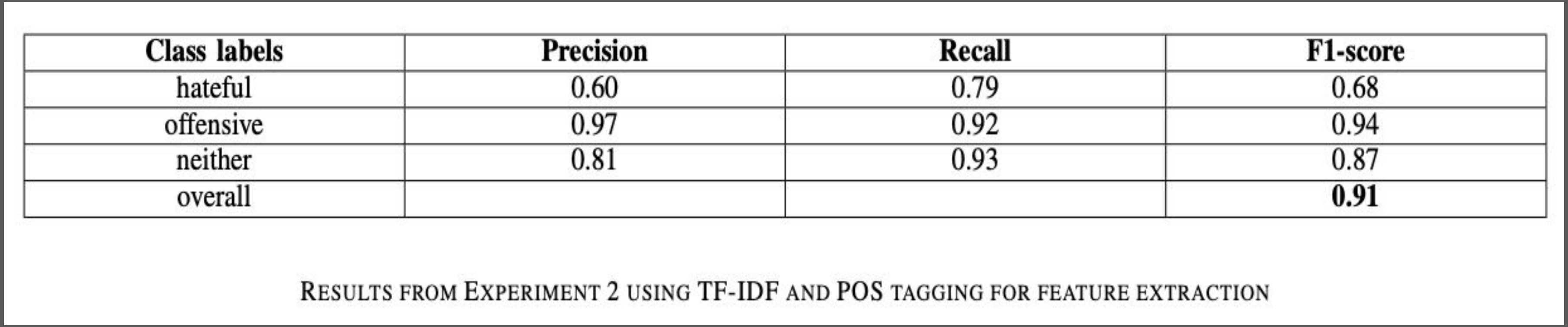
- L2 regularization
- Returns transformed version of input X
- This transformed input will be given as input to our classifier

# Model - Linear SVM

- Separate data points with a line across the hyperplane
- Goal is to maximize the margin and make a decision boundary
- One-versus-rest framework
- L2 regularization with 0.01 as regularization parameter
- Squared hinge loss function



# Results and Discussions



| Class labels | Precision | Recall | F1-score |
|---|---|---|---|
| hateful | 0.60 | 0.79 | 0.68 |
| offensive | 0.97 | 0.92 | 0.94 |
| neither | 0.81 | 0.93 | 0.87 |
| overall | | | **0.91** |

Results from Experiment 2 using TF-IDF and POS tagging for feature extraction

# Limitations

- Skewed dataset
- Hate comments without explicit toxic vocabulary
- Humorous content may be flagged due to use of controversial terms
- Overfitting can be addressed by using cross validation

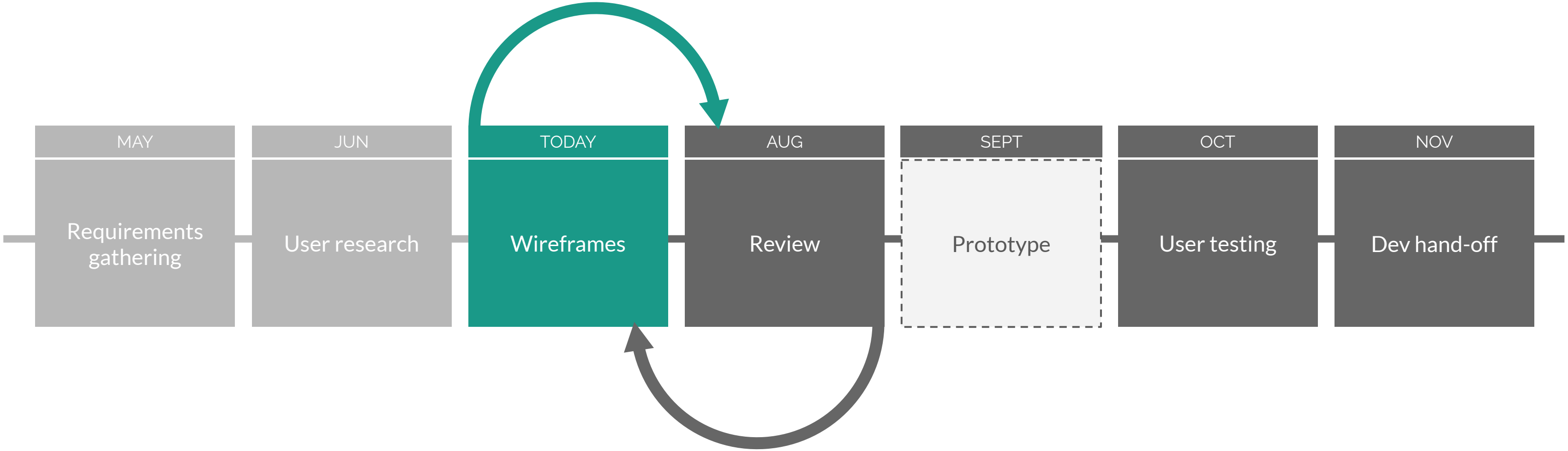
# Conclusion

# What next?

- Build richer dictionary
- Experiment with deep neural network architectures
- Use larger dataset for training

# Thank You!

# Timeline



| MAY | JUN | TODAY | AUG | SEPT | OCT | NOV |
|---|---|---|---|---|---|---|
| Requirements gathering | User research | Wireframes | Review | Prototype | User testing | Dev hand-off |