

# Masteroppgave

Bjørn-Andreas Lamo

Spring 2021

## 1 Introduction

Gender classification based on posted text on the internet has been in academic interest for decades[1]. And with emerging neural net models we are able to extract features from text that would otherwise be very difficult to detect [TODO].

single character to a 10 000 character long wall of text comment. The average comment length vary widely by subreddit. Some topics elicit long nuanced discussion, while others a funny quip. Most subreddits encourage consensus by sorting by most highly upvoted comments ascending. Others encourage debate by sorting by controversial first. Most users only look at the first

## 2 Author Profiling

## 3 Reddit

Reddit is a popular site to share content and discuss. The format is that users post links, photos/videos, or text, and other users can upvote or downvote the post based on how well it conforms with the *subreddit* it's posted in.

Subreddits is the way Reddit categorize different topics. The name scheme is r/SubredditName. For instance there is r/news for news in USA and r/worldnews for news worldwide excluding USA. These subreddits become communities as users regularly frequent them. Anyone can create their own subreddit to carve out a niche. This has led to the saying "There's a subreddit for that!".

Another important part of Reddit is the ability to comment on posts and reply to comments. The length of a comment can be as short as a

## 4 Methodology

## References

- [1] M. Koppel. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412, November 2002.