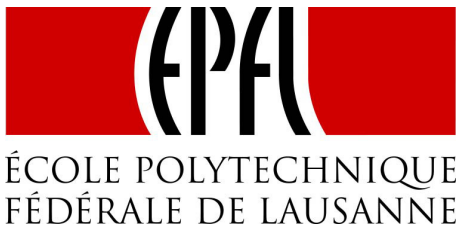
Twitter user's gender prediction

Gaétan Ramet Benjamin Schloesing Yuan Yao

A Network Tour of Data Science
Prof. Xavier Bresson
Prof. Pierre Vandergheynst
Michaël Defferrard

18th January 2016



Introduction

Objective: Gender prediction of twitter users

Steps :

- Data exploration
- Model training
- Data extraction
- Results

Data Exploration

Data Exploration

Data exploration

- Kaggle Dataset

	gender	description	link_color	profileimage	name	sidebar_color	text
0	male	i sing my own rhythm.	08C2C2	https://pbs.twimg.com/profile_images/414342229...	sheezy0	FFFFFF	Robbie E Responds To Critics After Win Against...
1	male	I'm the author of novels filled with family dr...	0084B4	https://pbs.twimg.com/profile_images/539604221...	DavdBurnett	C0DEED	ÛÏIt felt like they were my friends and I was...
2	male	louis whining and squealing and all	ABB8C2	https://pbs.twimg.com/profile_images/657330418...	lwtprettylaugh	C0DEED	i absolutely adore when louis starts the songs...
3	male	Mobile guy. 49ers, Shazam, Google, Kleiner Pe...	0084B4	https://pbs.twimg.com/profile_images/259703936...	douggarland	C0DEED	Hi @JordanSpieth - Looking at the url - do you...
4	female	Ricky Wilson The Best FRONTMAN/Kaiser Chiefs T...	3B94D9	https://pbs.twimg.com/profile_images/564094871...	WilfordGemma	0	Watching Neighbours on Sky+ catching up with t...
5	female	you don't know me.	F5ABB5	https://pbs.twimg.com/profile_images/656336865...	monroevicious	0	Ive seen people on the train with lamps, chair...

- 20k Tweets from different users, $\sim$ 5k with picture

Color features

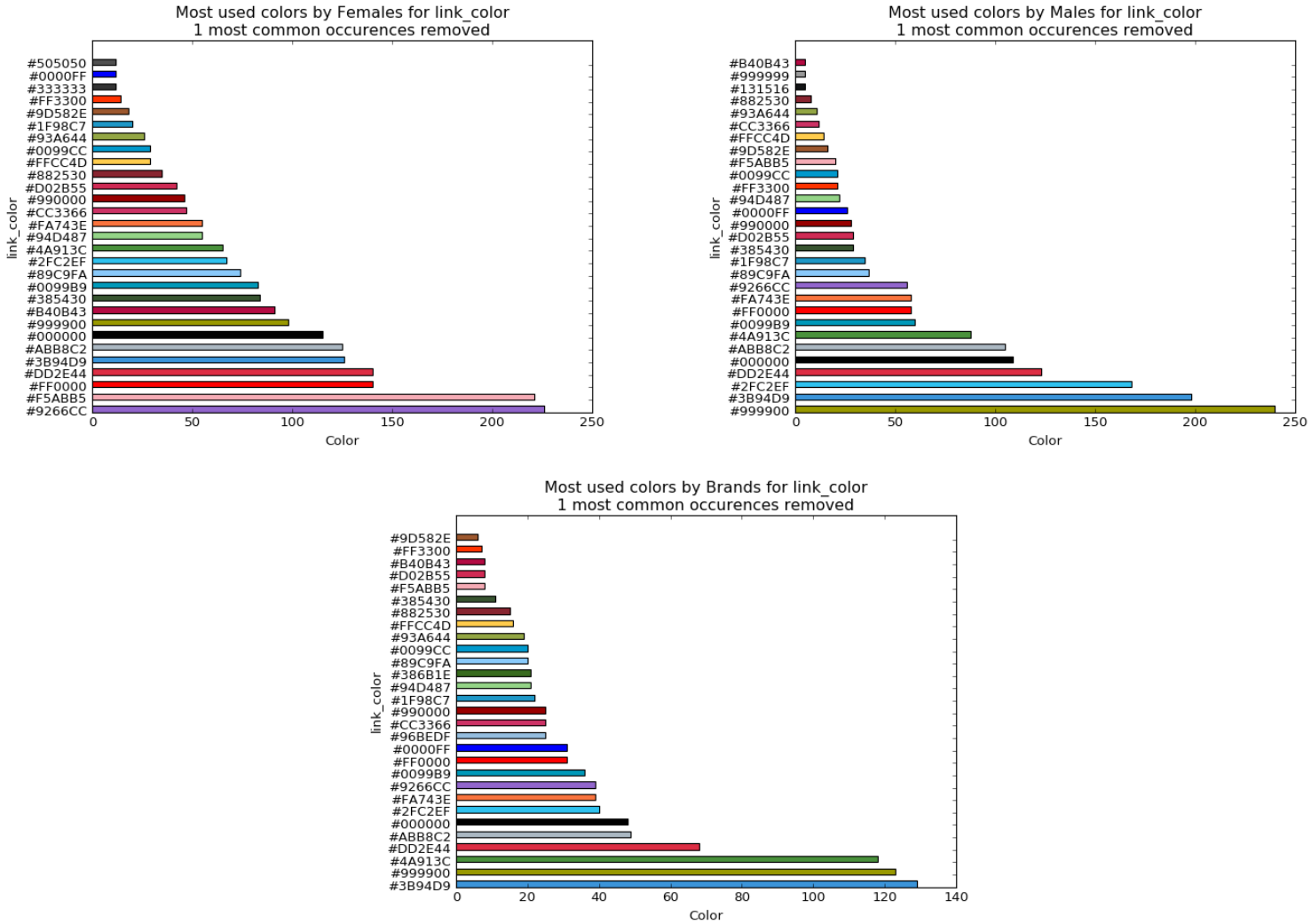


Figure 1: Most used link colors

Text features

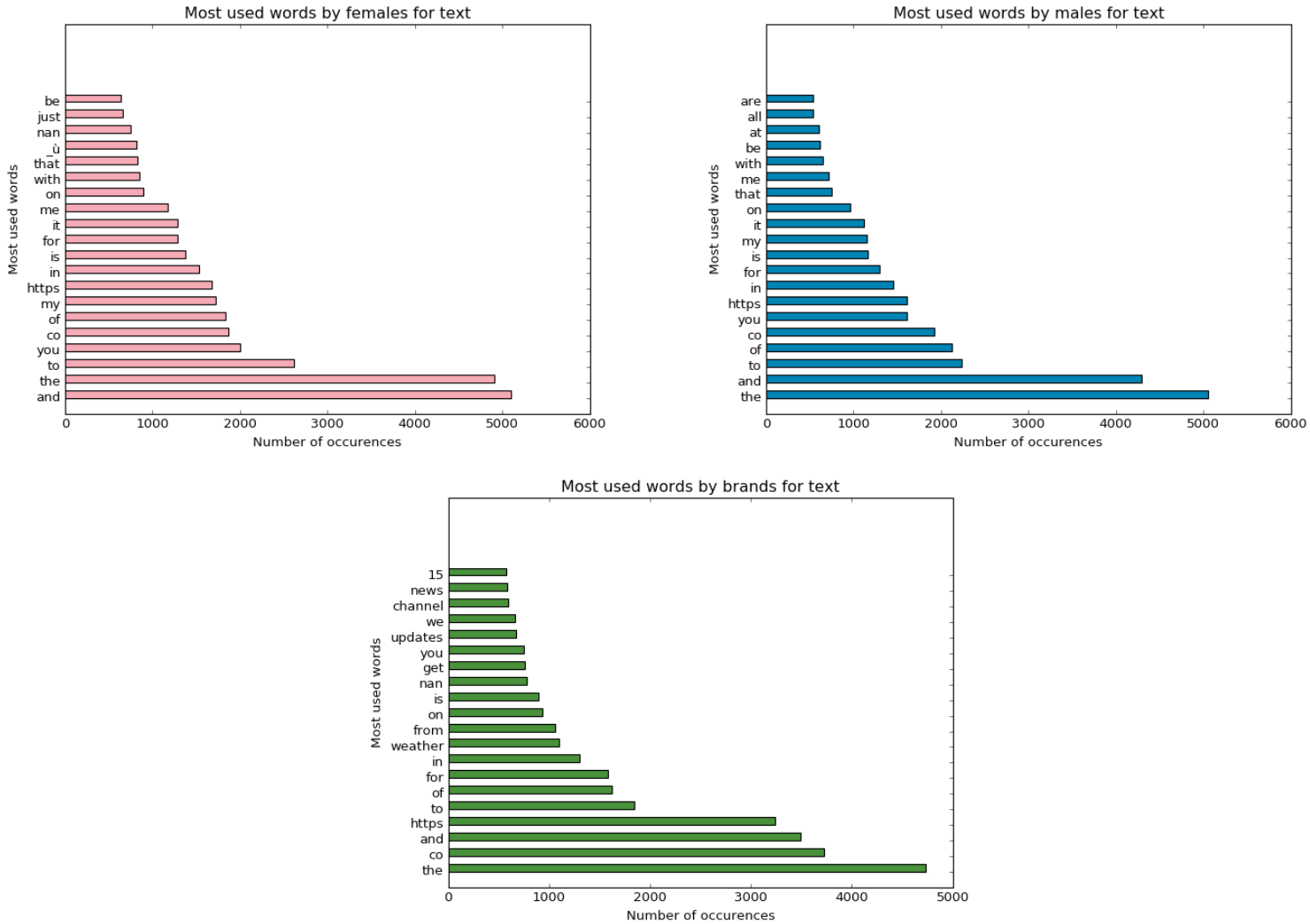


Figure 2: Most used words in tweets and users' descriptions

Profile pictures features

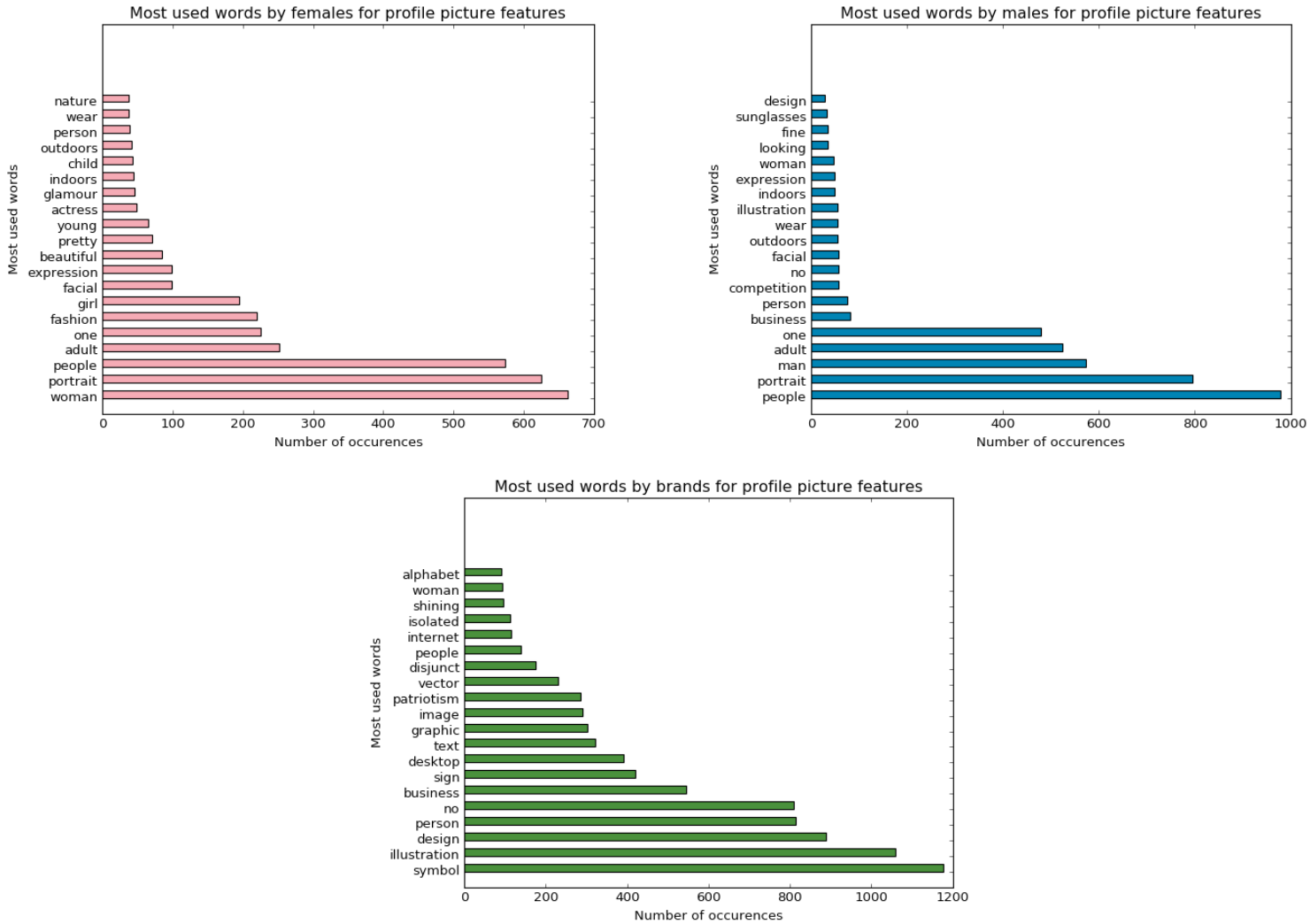


Figure 3: Most used profile picture contents

Model Training

Model Training

Model accuracy

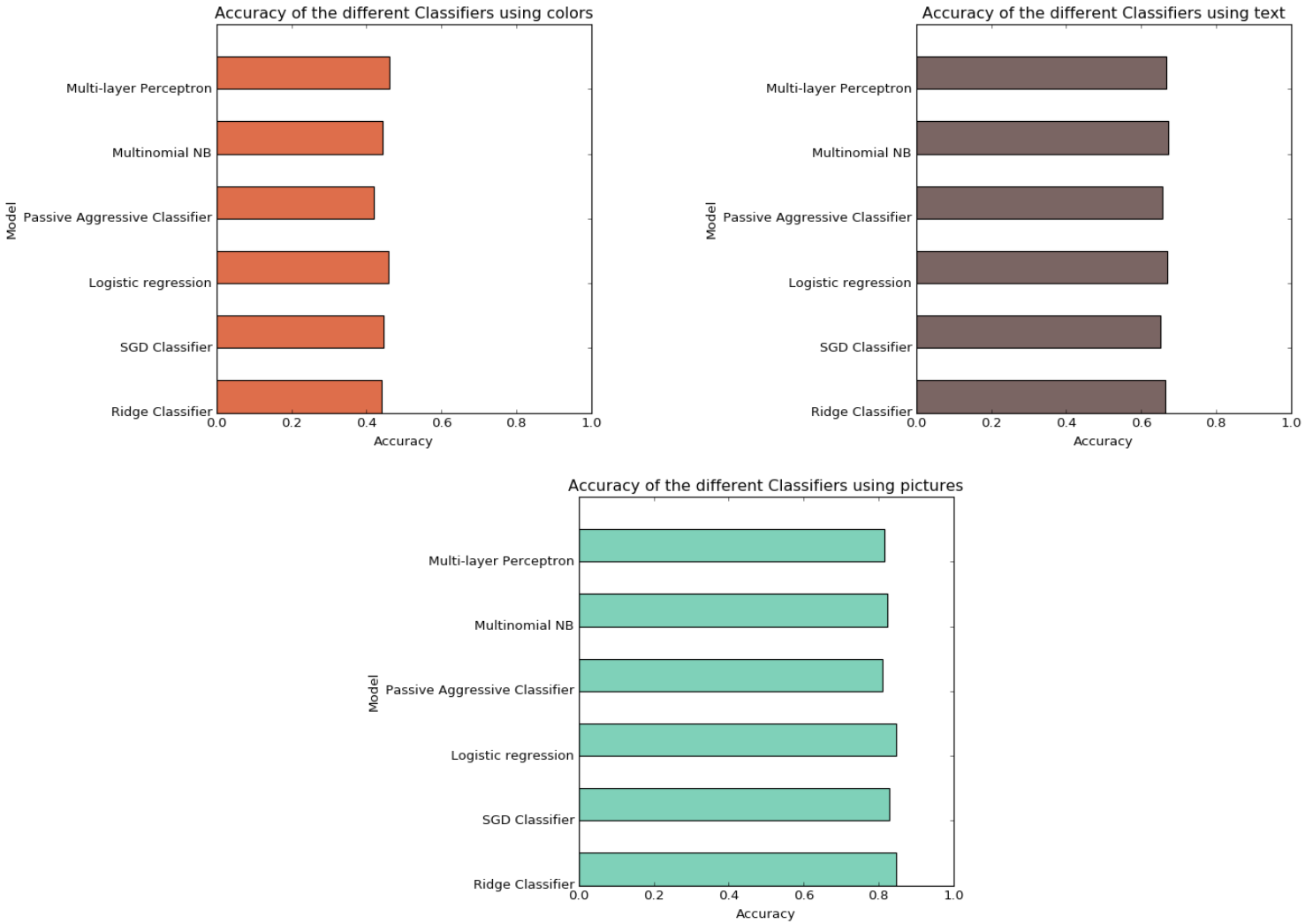


Figure 4: Accuracy of the models

Color Predictors

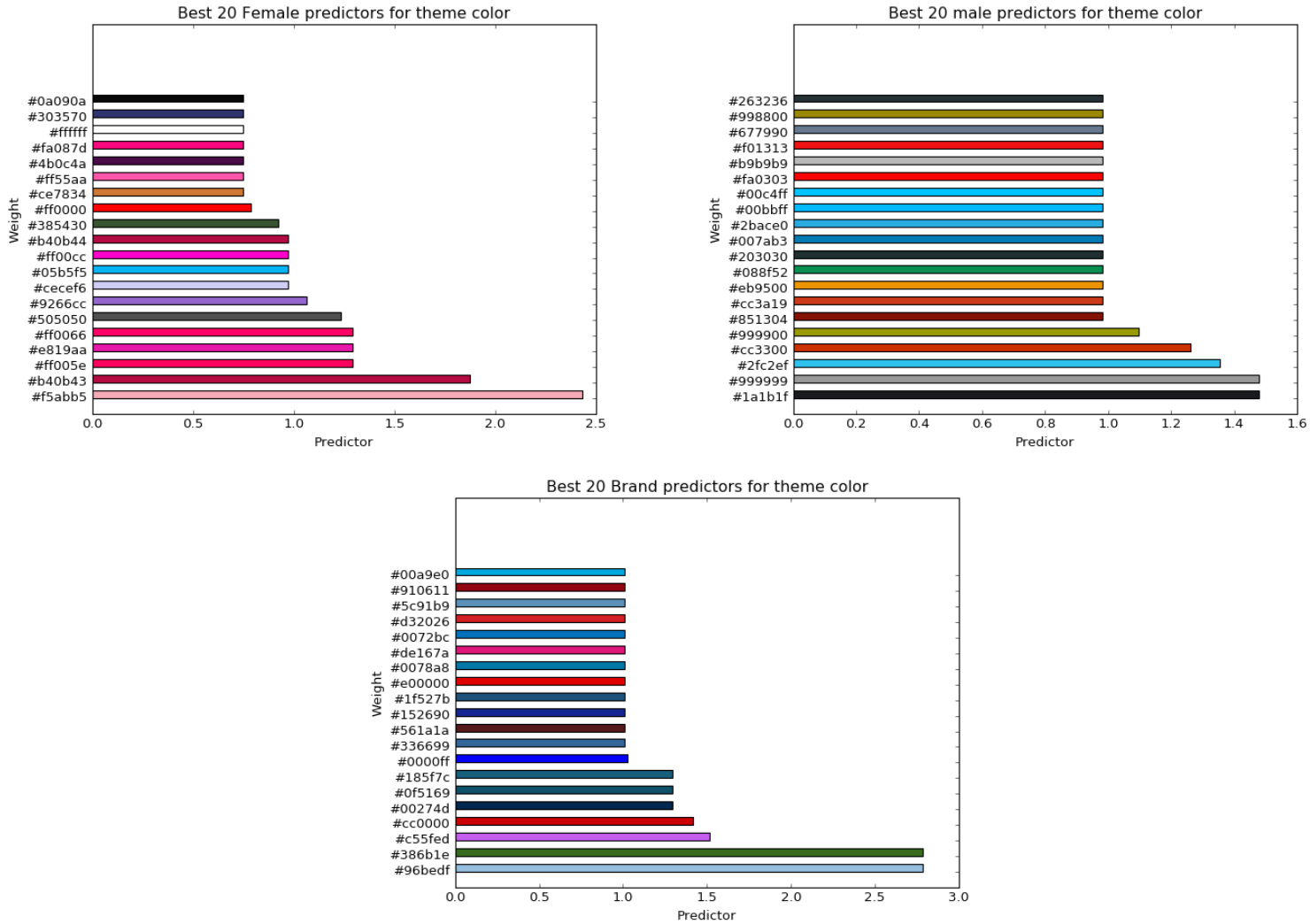


Figure 5: Predictors for color features

Color Anti-Predictors

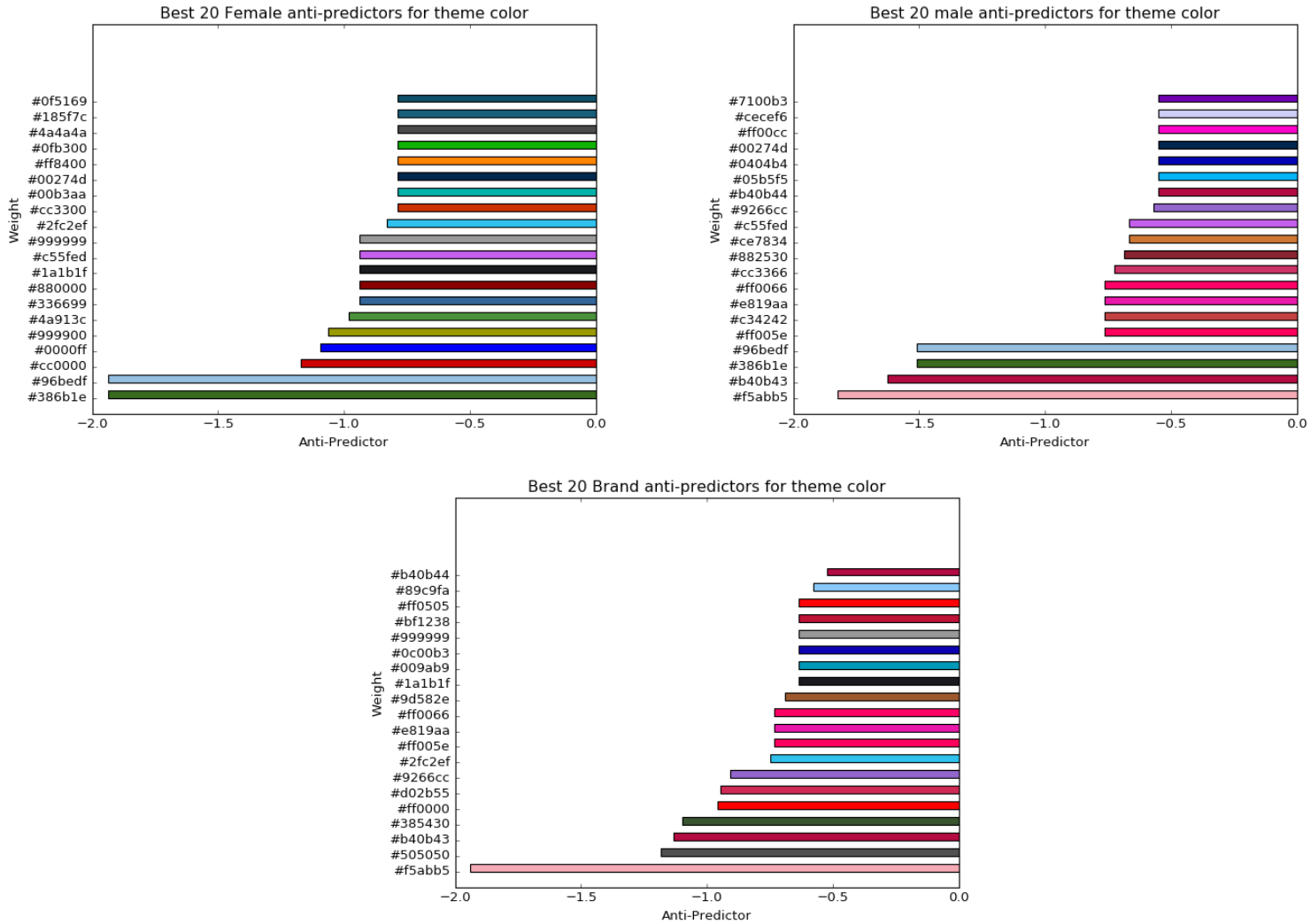


Figure 6: Antipredictors for color features

Text Predictors

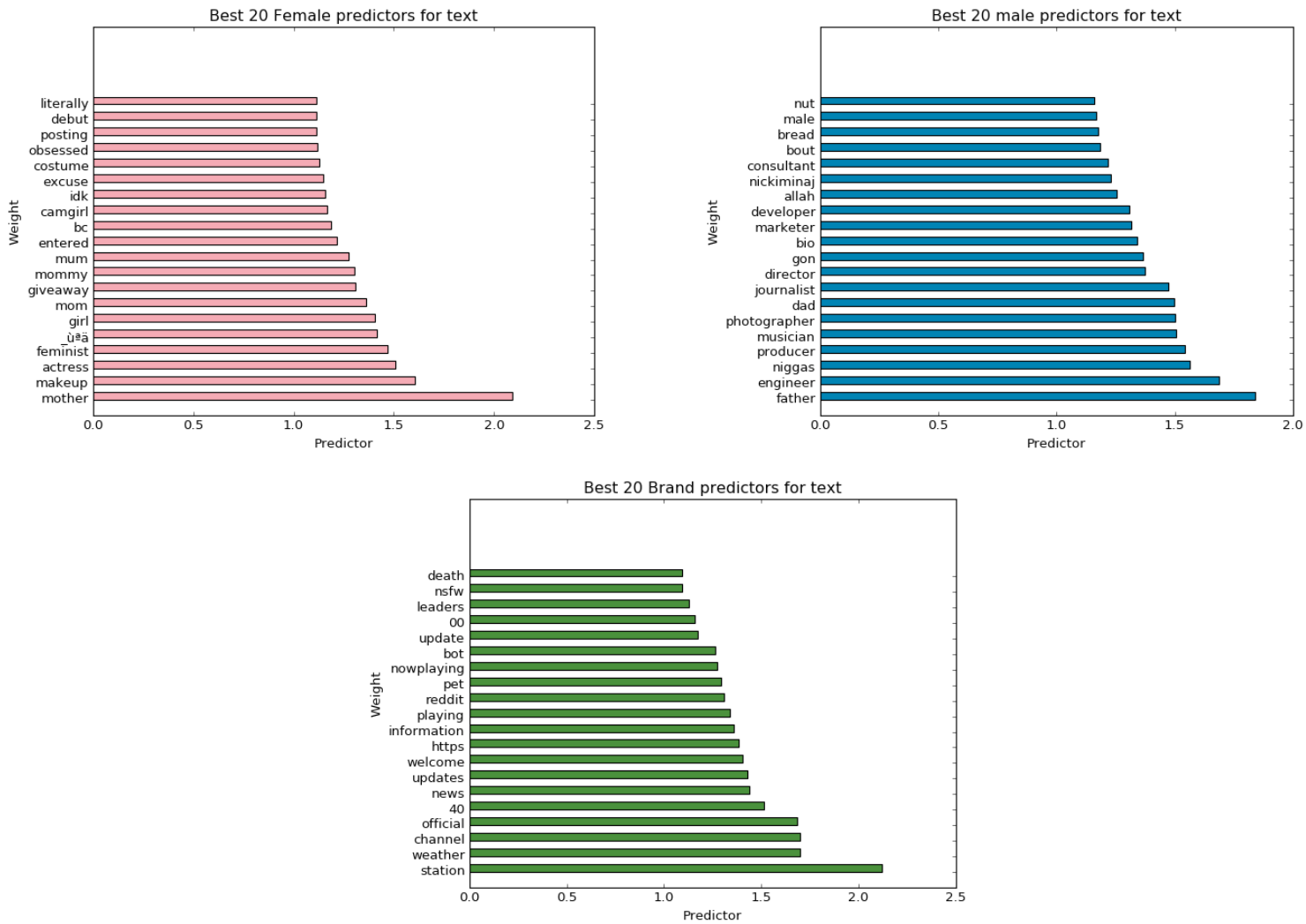


Figure 7: Predictors for text content

Text Anti-Predictors

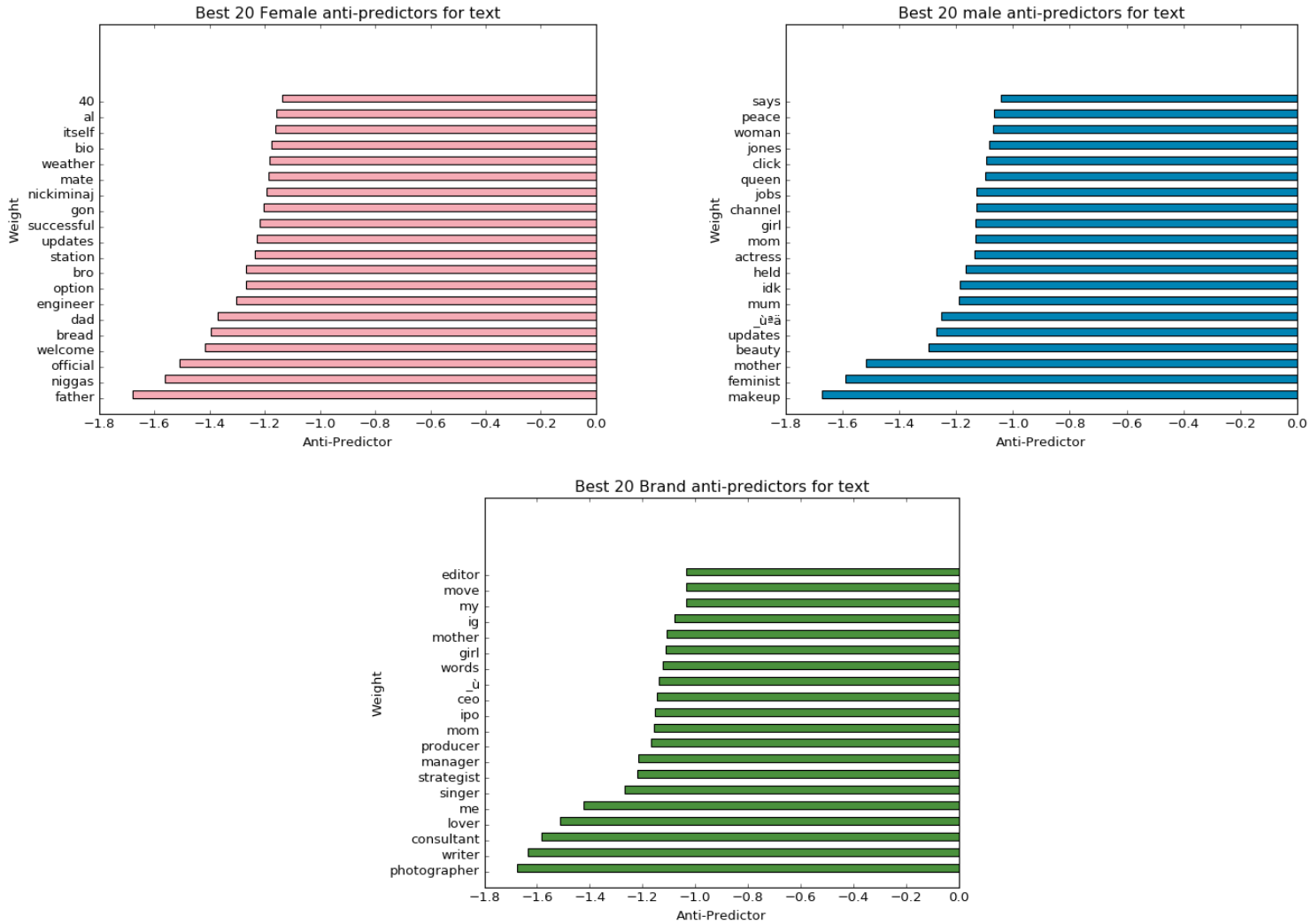


Figure 8: Antipredictors for text content

Profile picture features Predictors

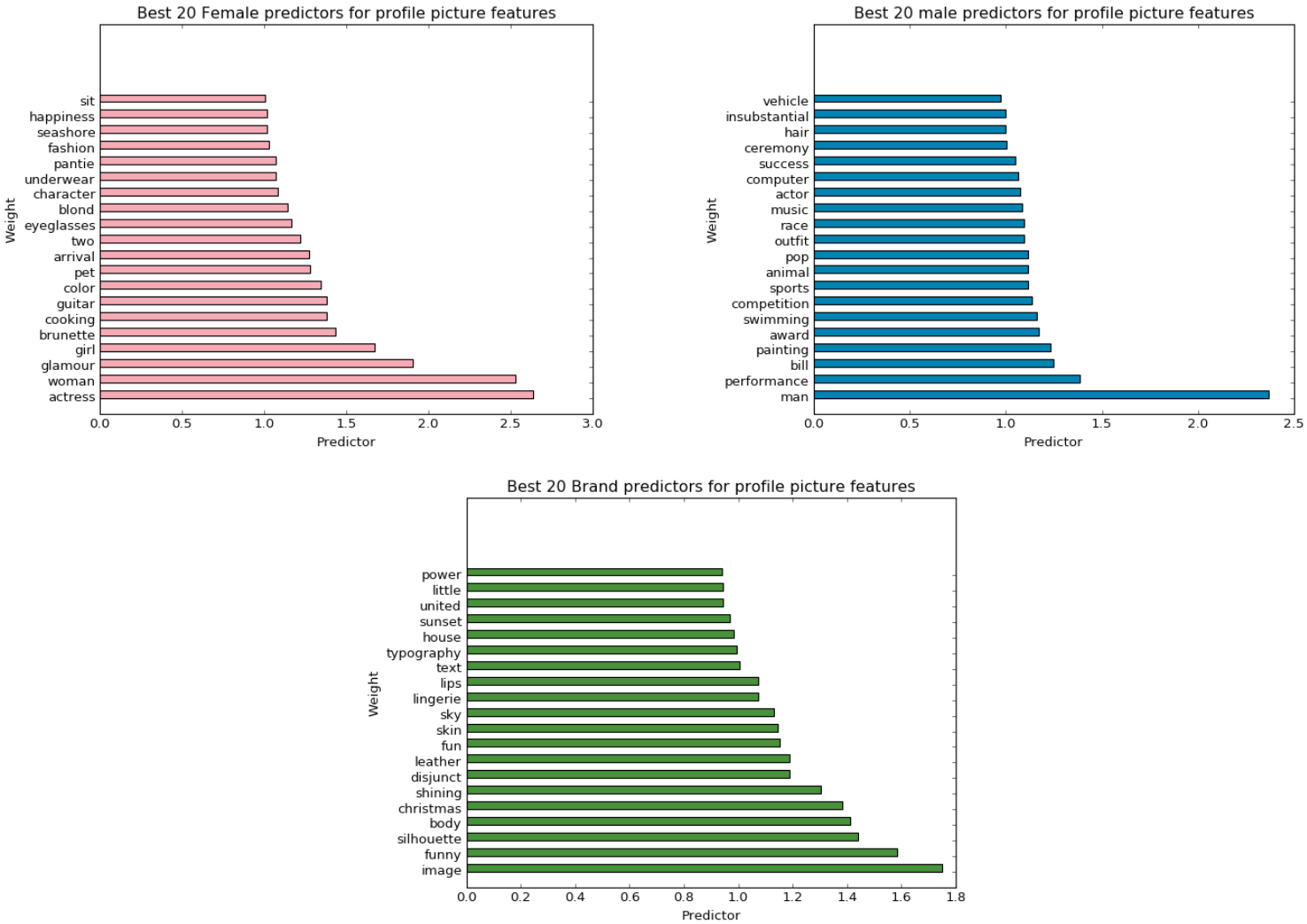


Figure 9: Predictors for profile picture features

Profile picture features Anti-Predictors

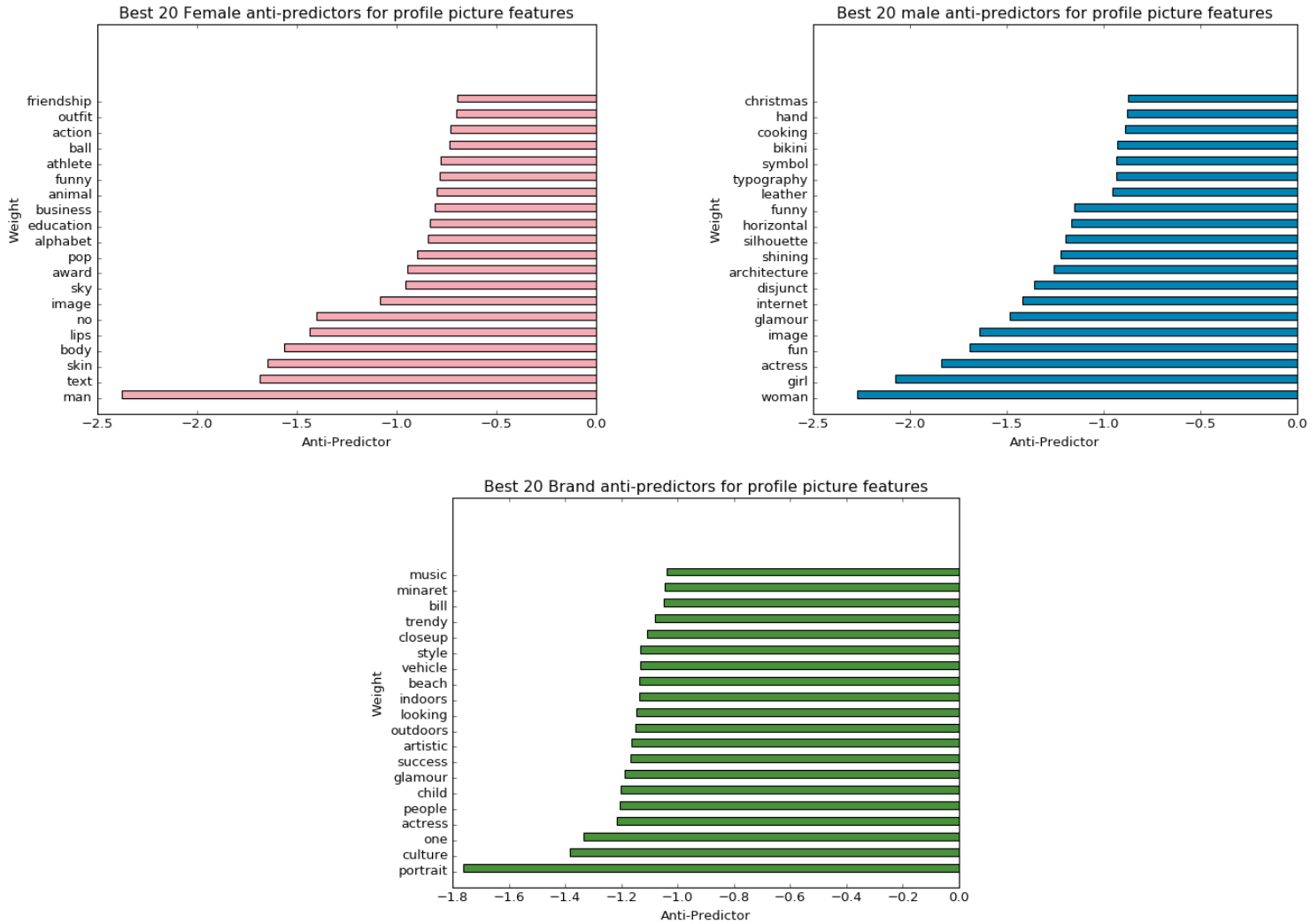


Figure 10: Antipredictors for profile picture features

Data Extraction

Data Extraction

Data Extraction

48 users : 16 from each class (brand, female, male) :

- Brands : Nestle, EasyJet, Deloitte, Toyota...
- Humorists, writers: Jimmy fallon, Paulo Coelho
- Singers: Sia, Rihanna
- Random people...

Dataset

	gender	description	link_color	profile_image_url	user_name	text
0	male	President-elect of the United States	6D5C18	http://pbs.twimg.com/profile_images/1980294624...	realDonaldTrump	INTELLIGENCE INSIDERS NOW CLAIM THE TRUMP DOS...
1	male	Author. Literary agent: @crsripley	000000	http://pbs.twimg.com/profile_images/7881514851...	augusten	RT @LUCYrk78: Only on page 12, and this is al...
2	male	NaN	000000	http://pbs.twimg.com/profile_images/3788000003...	Eminem	Until It All Falls Down, Have A Happy Holiday...
3	male	Read too much for my own good, write too much ...	C0DEED	http://pbs.twimg.com/profile_images/8201169243...	IrvineWelsh	RT @CreativeScots: Monday is the deadline for...
4	male	Books; articles: https://t.co/AOdtXKVYWQ; film...	C0DEED	http://pbs.twimg.com/profile_images/6445831904...	alaindebotton	RT @LeachJuice: I can't recommend this book e...
5	male	Writer of Submarine and Wild Abandon and poems...	F9FBFC	http://pbs.twimg.com/profile_images/6935333228...	joedunthorne	Just spilled “coffee” on my “manuscript”. (Wa...

Results

Results

Results

$$p(X = k) = \frac{\sum_{i=1}^{3} \alpha_i p_i(X=k)}{\sum_{i=1}^{3} \alpha_i}, \text{ with } \alpha_i = e^{10\text{acc}_i}$$

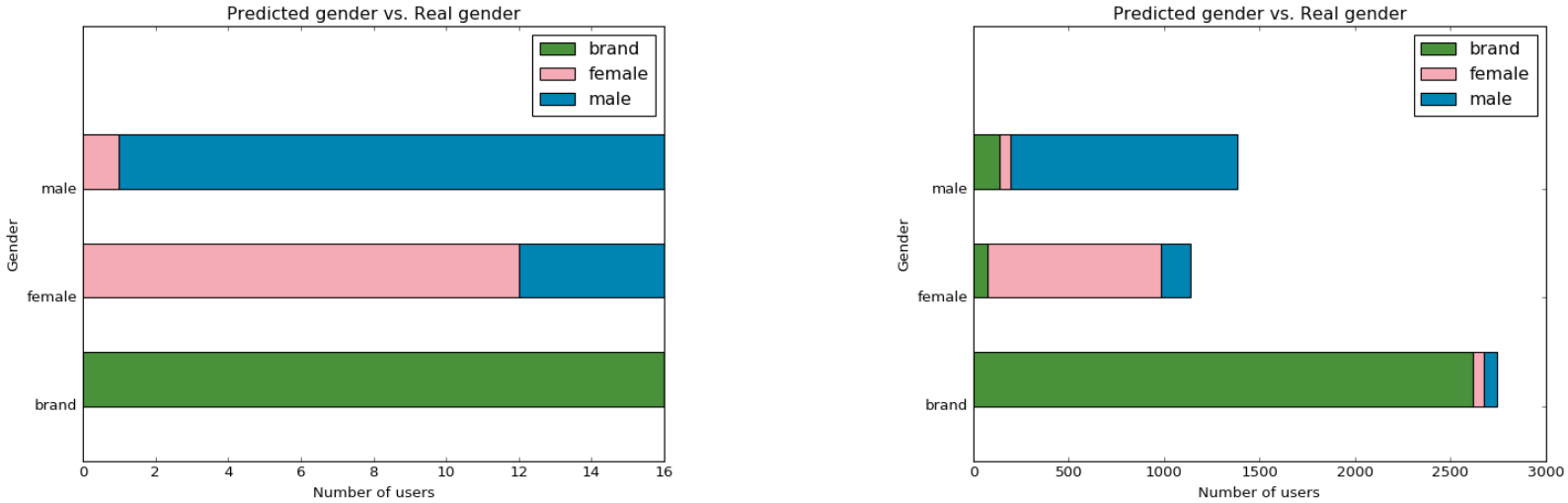


Figure 11: Testing the model on both the original dataset and the extracted one

Q & A

Thank you for your attention.