Gender information absolutely enhances sarcasm detection (/s)

Ruben Leonard van de Kerkhof
ANR: 320571

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COMMUNICATION AND INFORMATION SCIENCES,
MASTER TRACK DATA SCIENCE: BUSINESS & GOVERNANCE,
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

Thesis committee:

C.D. Emmery MSc
dr. Keuleers

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
December 2018

Preface

When I first started studying years ago, I did not expect to end up completing a master's degree in Data Science. There are a few people that have helped tremendously in completing the work that lies before you, as well as in my studies. I would like to thank my supervisor, Chris, for answering my many questions, excellent facilitation in terms of meetings and access to study materials, and, last but definitely not least, enthusing me about the many forms of machine learning approaches one could take for one single problem. I would like to thank my parents, brother and friends, for the continued support throughout the years, even when they were as surprised as I was when I decided to study Data Science.

Gender information absolutely enhances sarcasm detection (/s)

Ruben Leonard van de Kerkhof

In this work we investigate if there is gender information in sarcastic comments and subsequently if this can aid sarcasm detection. Most recent sarcasm detection research focuses on context and uses twitter as a data source, whereas we supplement the Self-Annotated Reddit Corpus (SARC) with distantly supervised gender labels gathered from Reddit. We evaluate gender prediction models for authors in the SARC which have a self-reported gender label after collection, and use the best performing model to predict gender labels on the rest of the SARC, after which gender is explicitly encoded in the comments. We conduct a pre-test to obtain an estimation of gender information in the SARC by predicting gender from sarcastic comments. Predicting gender from these comments proves difficult as the models do not perform above majority baseline. Sarcasm prediction from encoded comments does perform above baseline but no better than baselines from the SARC. Lastly, we provide detailed strategies for data analysis and collection on Reddit which could allow for future work to investigate other demographic characteristics on the SARC, or author profiling on Reddit in general.

1. Introduction

The widespread adoption of social media has caused the internet to become filled with a great amount of opinionated discourse across various social media platforms. Extracting information from these discussions can lead to new insights for business and research purposes. Several research fields have taken an interest in analyzing behaviour of people or other types of information on social media. Two research fields which are particularly interested in human behaviour on the internet are the fields of author profiling and sentiment analysis. Both of these fields find their application in the current study.

Author profiling concerns itself with inferring attributes of authors of texts, aiming to uncover identifying information such as gender or age from texts. Early work has shown there to be differences in language use and gender (Lakoff 1973; Pennebaker, Mehl, and Niederhoffer 2003). Applications of author profiling are found in marketing (understanding a company's demographic), forensics (understanding linguistic properties of scammers or harassers) and polling for political campaigns. Techniques that aim to predict gender from text have been used for some time, with Argamon et al. (2009) among the first. Understanding and identifying characteristics which can differentiate between gender can aid in inferring properties of authors by gathering salient details into features. On social media in particular, research has shown that there are differences in how men and women express themselves in certain contexts. For example; men more often refer to their significant other with possessive words than women do (Schwartz et al. 2013). Moreover, there is evidence that self-identified males and females on Facebook differ in tone, with women being more "people focused and warmer" and "men

being more socially distant" (Park et al. 2016). Lastly, research by Zheng et al. (2018) evidences that men and women differ in showing of emotions in online micro-blogs.

Where author profiling is concerned with identifying useful information about authors, sentiment analysis is concerned with understanding and analyzing opinions in text by using computational methods. Properly understanding opinions on various platforms can lead to new insights and or a better understanding of certain issues. Examples of these applications are, among others: understanding the reception of a product line, marketing campaign or the position of a company versus its competition. Apart from business applications, sentiment analysis can be useful for political campaigns and polling and recently has been used to investigate human trafficking (Mensikova and Mattmann 2018). The field of sentiment analysis has thusly enjoyed a great amount of attention in regards to research.

A major challenge in sentiment analysis is properly identifying sarcasm. Sarcasm, as defined in the Cambridge Dictionary, is "the use of remarks that clearly mean the opposite of what they say, made in order to hurt someone's feelings or to criticize something in a humorous way".[1] In other words, sarcasm can change the affective undertone of a piece of text. If a classifier is unable to pick up on sarcasm then the actual intended meaning of e.g. a product review may be misinterpreted or the results of a sentiment classifier may be influenced (Maynard and Greenwood 2014b). Therefore it is not surprising that sarcasm has become a big focus in sentiment analysis, as is evidenced by the article of Joshi, Bhattacharyya, and Carman (2017) in which they review research regarding automatic sarcasm detection.

Author profiling and sentiment analysis research have greatly benefited from social media data and platforms such as Twitter[2], evidenced by construction of corpora such as Twisty (Verhoeven, Daelemans, and Plank 2016) and Sentiment140 (Go, Bhayani, and Huang 2009). Therefore it is unsurprising that, due to the more easily accessible datasets concerning Twitter, most research regarding sentiment and sarcasm detection has been done on those types of datasets. With the recent publication of the Self-Annotated Reddit Corpus (SARC) by Khodak, Saunshi, and Vodrahalli (2018), containing over one million labelled comments as sarcastic or non sarcastic, and monthly Reddit comments provided by pushshift[3] (Baumgartner[4]), author profiling and sentiment analysis can extend to different social media platforms.

Social media platform Reddit[5] differs from Twitter in a number of ways which can influence language use. Most notable of which is that Reddit posts are not limited to 140 characters and offers subcommunities called subreddits in which users gather to discuss certain subreddit specific topics. Like Twitter with 'RT' ('Retweet'), Reddit has unique Reddit-only language. An example of this is that subreddits can have specific words which illustrate certain common ideas or aspects used in that subreddit, such as 'AMA' which stands for 'Ask me anything', 'TIL' ('Today I learned'), 'FTFY' ('Fixed that for you') or 'RR' ('Recommended routine'). Some of these terms are used across subreddits, of which 'AMA' and 'TIL' are both examples. A big difference between Twitter and Reddit, however, is that Reddit has more informal language use as there are fewer companies and public institutions represented on Reddit. A final significant distinction between platforms is that Twitter has more easily identifiable authors since

1 https://dictionary.cambridge.org/dictionary/english/sarcasm
2 https://twitter.com/
3 https://pushshift.io/
4 https://pushshift.io/what-is-pushshift-io/
5 https://www.reddit.com/

they can provide their Twitter profile with demographic information and pictures of themselves. In contrast, Reddit does not have such a detailed profile page on which users can provide demographic information or pictures (even though users can provide this in comments on threads) and it is generally not recommended to provide such information on the platform. In other words, Reddit is more anonymous than Twitter.

Although these differences can make it more difficult to conduct research on Reddit, the platform has been gaining popularity for data analysis due to initiatives such as PRAW[6] and the earlier mentioned pushshift. This is illustrated by the work of Wallace, Choe, and Charniak (2015), in which they investigated 'verbal irony' (a term used by the authors to refer to a mixture of sarcasm and irony) by using certain specific subreddits as a surrogate for user communities in order to create context and combined this with sentiment and extracted noun phrases to improve precision at no cost to accuracy when detecting verbal irony. While Hazarika et al. (2018) investigated sarcasm classification on the SARC by using user embeddings and post context.

As mentioned earlier, there are apparent differences in writing styles between gender and how men and women express themselves online (Schwartz et al. 2013; Park et al. 2016; Zheng et al. 2018). Research also shows that there gender differences in sarcasm, with men more frequently using sarcastic remarks (Dress et al. 2008) and differences in how men and women respond emotionally to sarcastic remarks (Drucker et al. 2014). Even though much research has been done regarding gender classification and sentiment analysis, and (as mentioned above) research has shown that there are gender differences in sarcasm usage and writing styles, little research has been done regarding demographic and stylometric differences in usage of sarcasm in social media text. A possible explanation for the limited research on stylistic or linguistic differences in sarcasm across demographic variables is that on social media websites and discussion forums it is usually not required to provide personal information such as age or gender which means that for supervised learning applications labels with demographic information will be missing. Research by Emmery, Chrupała, and Daelemans (2017) has shown that through simple queries, distant labels are able to be obtained for demographic information regarding social media users.

The current study aims to study gender difference in sarcasm use by using the recently published SARC and data from the pushshift initiative. There are two main research questions of which the first is: "Can machine learning methods detect gender information from sarcastic messages?", and the second is: "Can gender information be used to enhance sarcasm classification?". The first research question will be answered by predicting gender from sarcastic and non-sarcastic messages by using the SARC supplemented with distantly gathered labels, using a data collection strategy inspired by Emmery, Chrupała, and Daelemans (2017). Stylometric features will be used to predict gender from these comments, and the expectation is that it is possible to effectively predict gender from these messages. The first research question functions as a pre-test, and allows us to obtain an indication of the amount of gender differentiating information in the comments and authors in the SARC, and if this is sufficient enough to effectively classify gender. This will then be treated as the basis to answer the second research question, by using the best performing model to predict gender labels. This means that the reliability of our predicted gender labels on the remainder of the SARC depends on the results of the pre-test, and can therefore influence the results of our second research question. The pre-test therefore also provides information towards understanding the

6 https://praw.readthedocs.io/en/latest/

reliability and validity of the results of the second research question. The second research question "Can gender information be used to enhance sarcasm classification?" will be answered by predicting sarcasm after explicitly encoding gender information into sarcastic and non-sarcastic messages by using the strategy outlined in Johnson et al. (2016), their strategy will be discussed more in the experimental procedure. If encoding gender information improves sarcasm classification performance above baselines given by Khodak, Saunshi, and Vodrahalli (2018) then gender could be a important variable to consider in sarcasm detection.

The contributions of the current study pertain to both the field of author profiling and sarcasm detection, as probably one of the first studies to combine aspects of both fields to investigate if gender information can enhance sarcasm detection. Furthermore, we add to the body of research conducted on Reddit by using distant supervision to obtain salient information. Moreover, we add to research done on the SARC by taking a different direction than most research that has used the SARC, which is the use of context to enhance sarcasm detection. Instead, we opt to explore if gender information can be used for sarcasm detection on Reddit. We provide a detailed data collection approach concerning Reddit and different approaches to ensure quality labels for distant supervision as well as pre-processing steps and evaluate different train and test strategies.

2. Related Work

2.1 Sentiment Analysis

Sentiment analysis emerged from the fields of automatic text and topic categorization to structure online information. One of the earliest to investigate automatic sentiment classification were Pang, Lee, and Vaithyanathan (2002) focusing on using machine learning algorithms to structure documents by its predicted sentiment instead of its topic. They acknowledge that the techniques used in their research based on bag-of-features have difficulty with recognizing a deliberate contrast in the text, such as "This is horrible, I love it". In other words, there is a change in the meaning of a sentence which can influence the correct classification. As stated earlier, a somewhat equivalent situation arises with sarcasm, since sarcasm can change the actual sentiment of a sentence or piece of text as well. While Pang, Lee, and Vaithyanathan (2002) indicate the need to use techniques which can detect signals of whether sentences with a deliberate contrast are close to the overall topic described in the piece of text, others have focused on detecting other linguistic devices which can change the polarity of a sentence and therefore influence sentiment detection, such as sarcasm. Maynard and Greenwood (2014a) have demonstrated the impact sarcasm has on sentiment detection on Twitter, and have therefore touched on the importance of sarcasm detection.

2.2 Sarcasm detection

Sarcasm detection is a subfield of sentiment analysis and one of the earliest to investigate were Tepperman, Traum, and Narayanan (2006) who focused on sarcasm in speech with decision trees. Sarcasm and verbal irony are often used interchangeably in research (Kreuz and Glucksberg 1989; Wallace, Choe, and Charniak 2015), whereas Joshi, Bhattacharyya, and Carman (2017) acknowledges sarcasm as a specific type of verbal irony designed to scorn. Sarcasm has been researched from different perspectives such as the linguistic perspective, which led to different theories on sarcasm (Kreuz and

Glucksberg 1989). In the current study, however, the focus will be on automatic sarcasm detection by using computational and machine learning techniques.

Research in automatic and computational sarcasm detection can be divided into different categories such as: rule-based, (semi-)supervised, statistical (with features such as average word length, etc.) and deep learning (Bouazizi and Ohtsuki 2016; Joshi, Bhattacharyya, and Carman 2017; Raghav and Kumar 2017). An example of rule-based sarcasm detection is the work of Maynard and Greenwood (2014b) who used hash-tag based patterns on Twitter to create rules, while others were more focused on patterns concerning emoticons and punctuation (Read 2005; Carvalho et al. 2009). Recent trends in sarcasm detection have been to use context to detect sarcasm (Wallace, Choe, and Charniak 2015; Wang et al. 2015; Hazarika et al. 2018) and distant supervision to collect indicators of sarcasm which is primarily applied on Twitter through the use of hash-tags (Reyes, Rosso, and Buscaldi 2012; Liebrecht, Kunneman, and van Den Bosch 2013). Most research has been conducted on Twitter data, however, Khodak, Saunshi, and Vodrahalli (2018) used the distant supervision approach to construct a dataset based on Reddit, which will be used in the current study.

Supervised sarcasm detection approaches use a variety of features among which are surface pattern features such as high frequency words and content words (Davidov, Tsur, and Rappoport 2010), punctuation and word-embeddings (Ghosh, Guo, and Muresan 2015), syntactic features such as n-grams and Bag-Of-Words (BOW) (Gonzalez-Ibanez, Muresan, and Wacholder 2011) and Part-Of-Speech tags (POS) (Bouazizi and Ohtsuki 2016). In terms of supervised classification algorithms, just as in sentiment analysis, support vector machines (SVM), logistic regression and Naive Bayes are among the most popular due to their effectiveness in text classification (Abbasi, Chen, and Salem; Raghav and Kumar 2017), with numerous research making use of these techniques.

Even though sarcasm detection has recently enjoyed a lot of attention, there has not been an investigation into whether gender information can enhance sarcasm classification. Research has shown that there are differences in how men and women 'enjoy' sarcasm (Drucker et al. 2014), that there are regional differences for gender in frequency of sarcasm usage where men more often (based on self-reports) used sarcasm (Dress et al. 2008), and that (based on self-assessment) men overall were more sarcastic and women were more sarcastic when talking to men (Rockwell and Theriot 2001). However, this information has not yet been used in sarcasm detection work, which can be due to several reasons, for example due to the lack of annotated data. Gathering labels for gender information can be a time-consuming task, this could be remedied by using distant labels in the same manner distant supervision approaches have done in investigating sarcasm. A brief overview of distant supervision is given in subsection 2.5.

To summarize, sarcasm detection research has focused on context, distant supervision, patterns and rule-based systems. In terms of data it has mostly been focused on analyzing Twitter datasets. Lastly, there has not been a focus on gender information in sarcastic (online) text. Which is why the current study will be focused on Reddit data and employ strategies based on distant supervision and from the field of author profiling to explore if supplying gender information can enhance sarcasm classification. This study will fall under the supervised classification category of sarcasm detection, treating sarcasm as a binary classification problem.

2.3 Self-Annotated Reddit Corpus

As mentioned earlier, the current study will use the SARC by Khodak, Saunshi, and Vodrahalli (2018) since it is a relatively new and unique sarcasm corpus in the sense

that it contains over 1 million comments and is one of the few sarcastic corpora based on Reddit. Due to its recent publication, much research has not been done using the SARC. There have been a few works with some interesting approaches. Hazarika et al. (2018) found drastic improvements in classification using neural networks and extracting context information aside from user-embeddings. Kolchinski and Potts (2018) opted to estimate a user's tendency to be sarcastic and used that in combination with dense embeddings in a recurrent neural network. Most of the approaches based on the SARC have sought to model context to improve classification. Khodak, Saunshi, and Vodrahalli (2018) have conducted a number of baseline tests, using linear models with features such as bag-of-words, bag-of-bigrams and sentence embeddings as well as models which used context. When evaluated on balanced parts of the corpus the simple linear classifiers perform above baseline, albeit lower than human evaluation. In the unbalanced settings the performance suffered greatly. Most of the approaches concerning the SARC have used neural networks and in some form attempted to model context, whereas the current study will use a supervised approach and use data gathered from Reddit to supplement the SARC with gender information.

2.4 Author Profiling

Where automatic sarcasm detection is a subfield of sentiment analysis, author profiling is a subfield of computational stylometry, with early work by Mosteller and Wallace (1963). Computational stylometry can be divided into three subfields, authorship attribution, identification, and profiling. Attribution and identification deal with uncovering the author of a text, whereas profiling is interested in uncovering, amongst others, specific demographic characteristics of authors. For the latter, common interesting characteristics are age and gender. Early work showed that there are differences in how men and women use language (Lakoff 1973; L. Newman et al. 2008; Schwartz et al. 2013; Park et al. 2016), and the work of Koppel, Argamon, and Shimoni (2002) is widely regarded as one of the first to show that demographic characteristics can be predicted from text.

The field of author profiling has, with the advent of online text and social media, made good use of the increase in data, with researchers making use of blogs (Schler et al. 2006; Argamon et al. 2009) as well as e-mails (Estival et al. 2007). More recently, social media websites have gathered interest from the research community and, just as in sentiment analysis, Twitter has been a predominant source of data and area of interest (Verhoeven, Daelemans, and Plank 2016; Rangel et al. 2017, 2018). Most work in author profiling use a supervised approach, and to gather information to predict demographics in a supervised setting, labels are needed. As such, just as in sarcasm and sentiment detection, when investigating profiling on social media, a distant supervision approach has been used to gather information for labels (Li, Ritter, and Hovy 2014; Culotta, Ravi, and Cutler 2016; Emmery, Chrupała, and Daelemans 2017).

Features used can be divided into lexical, syntactic, semantic, structural and domain specific features (Neal et al. 2017). Commonly used features are features such as n-grams of characters and words, such as in Ferreira and Neto (2017), where they combined style features with n-grams to predict the gender of Twitter users. Sanchez-Perez et al. (2017) compared various character n-gram levels for gender prediction and observed that larger n-gram values had the best accuracy in classifying author gender in Spanish texts, whereas Sarawgi, Gajulapalli, and Choi (2011) found that lower level character n-grams performed better and were more robust against topic influence. Sboev et al. (2016) also investigated gender with a focus on topic independent features but on Russian texts, and observed that their deep learning method reached high accuracy. Simaki, Mporas,

and Megalooikonomou (2016) evaluated the importance of several features used in their prediction of gender and age on social media and concluded that pronouns, articles and adverbs are among the best distinguishing features for gender detection. Discourse and syntactic features have also been used to identify gender by (Soler Company 2016), in which, among others, they used POS-tags, dependency trees and features based on the shape of the dependency tree as well several discourse relations. They found that syntactic, which were the POS-tags and dependency trees, and character based models worked best on their own, and all features together had the best accuracy.

To summarize, author profiling is concerned with inferring characteristics of authors by using a variety of text sources as data. The success of distant supervision and noisy labels has allowed the field to leverage social media as a data source. Profiling is mostly approached from a supervised classification perspective, making use representations such as bag-of-words and features such as n-grams or syntactic and discourse features.

2.5 Distant supervision

As briefly mentioned earlier, distant supervision has been used as a strategy to obtain labels for supervised classification approaches in both sarcasm detection and author profiling, it also provides an effective strategy for the otherwise arduous task of manual annotation. The idea behind distant supervision for author profiling and sentiment analysis is closely related to that of relation extraction (Mintz et al. 2009) in the sense that; if an author at some moment states a characteristic about itself, then we can assume other texts produced by this author to carry information regarding said characteristic. This technique has been used to obtain labels whether a message is sarcastic or not (Reyes, Rosso, and Buscaldi 2012; Liebrecht, Kunneman, and van Den Bosch 2013) or to gather demographic information (Li, Ritter, and Hovy 2014; Culotta, Ravi, and Cutler 2016; Emmery, Chrupała, and Daelemans 2017).

Even though this allows researchers to use a wider array of data to conduct experiments, it is not without risks. Collected labels can cause a skew or a bias in the data (van Dalen, Melein, and Plank 2017). An example of their collection strategy returning biases is that they found that oftentimes twitter users that stated their political alignment are politicians themselves. This 'danger' of noisy labels has also been mentioned by Bamman and Smith (2015) for noise in Twitter hash-tags of sarcasm, as well as by Khodak, Saunshi, and Vodrahalli (2018) in the creation of the SARC. The latter of which have taken steps during their corpus creation to combat this by manual inspection. To reduce noisiness they, among other strategies, manually investigated false positives and false negatives from a random sample of both sarcastic and non-sarcastic comments which their corpus creating technique detected. To determine if a post was sarcastic its post-context (the comments leading up to sarcastic comment) was provided to human annotators. Furthermore, they examined the way the sarcasm tag (/s) was used in the comment. They removed comments which refer to the tag, or use the tag in the middle of the sentence. Lastly, they acknowledge that in an unbalanced setting the ratio of false negatives to sarcastic comments is quite high. In other words, the research by Khodak, Saunshi, and Vodrahalli (2018); van Dalen, Melein, and Plank (2017) show that effort is needed to obtain quality labels through distant labeling. Therefore, we include detailed steps to obtain accurate labels in the current study.

2.6 Current study

In the current study we will explore if there is gender information in sarcasm by using a pre-test which entails predicting gender from sarcastic Reddit comments, provided by the SARC (Khodak, Saunshi, and Vodrahalli 2018). To obtain gender labels, we will apply a data collection strategy inspired by the simple query method from Emmery, Chrupała, and Daelemans (2017) to several months of Reddit comments, provided by the pushshift[7] initiative. A fastText[8] classifier will be used as a baseline with character and word n-grams since it has proven to be an efficient and quick text classifier (Joulin et al. 2016). POS-tags will be used to see if it can improve gender classification. Afterwards, gender information will be captured in the comments following the strategy by Johnson et al. (2016) and used to train and test models in a sarcasm prediction task.

3. Experimental Setup

The current study aims to explore if machine learning methods can find gender differentiating information in sarcastic messages. Furthermore, it is investigated if encoding gender information in features can predict sarcasm on Reddit. This is attempted by predicting gender from comments in the Self-Annotated Reddit Corpus (SARC) and predicting sarcasm from non-sarcastic and sarcastic comments. Both research questions are treated as a binary classification problem.

The current study uses four datasets, one is the SARC created by Khodak, Saunshi, and Vodrahalli (2018) and acquired from Kaggle[9]. The other data sets are collected from Reddit comments in the period of May 2018 until September 2018, which were made available on a MongoDB server. First the SARC will be discussed, afterwards the collection process to obtain gender labels from the monthly Reddit comments will be discussed, thereafter the pre-processing steps taken and as last the experiments.

3.1 Data

3.2 Self-Annotated Reddit Corpus

The SARC was created by Khodak, Saunshi, and Vodrahalli (2018) and the version that is used in the current study contains 1,010,826 Reddit comments labeled as either sarcastic (1) or not sarcastic (0). It is a balanced data set, which contains equal amounts of sarcastic as non-sarcastic comments. Comments were gathered by the authors by searching for a sarcasm tag, denoted by /s, and span Reddit comments from 2009 until April 2017. Note that the /s tag is not present in the data set comments. The authors also removed URLs. The SARC also contains the author names belonging to the comment, and each author has at least one sarcastic and not-sarcastic comment. Some comments in this data set contain mentions to subreddits (which are sub communities on Reddit) or other usernames. The dataset used in this study does not contain parent comments or the comments leading up to the sarcastic or non-sarcastic remark as context is not investigated.

7 https://pushshift.io/
8 https://fasttext.cc/
9 https://www.kaggle.com/danofer/sarcasm

3.3 Gender-label collection

To obtain gender indicators, a search strategy inspired by Emmery, Chrupała, and Daelemans (2017) was employed, but adapted to Reddit comments and structure. Reddit allows users and moderators (caretakers of a specific subreddit) to set flairs. Flairs are a way of adding information to a user for a specific subreddit, which can mean that flairs can differ across subreddits.

There are two types of flairs in the data structure of the pushshift comments, "author flair text" and "author flair css class". A simple example would be a flair text with: 'F' which could stand for female, or a css class with 'female'. Author flair text can be set by users themselves, whereas the css class can be set by moderators. Therefore, to obtain gender labels, author flair text and css class were queried against the following words: M/male, F/female, M/m and F/f. To obtain gender labels for authors in the SARC, the set of unique authors from that dataset were obtained, and subsequently labels were sought, as well as for authors which do not appear in the SARC. Note that for the different sets of authors (SARC and non-SARC) it was ensured that no author appeared in both sets.

To inspect the reliability of the obtained labels, the results were sorted by frequency of labels and for the top 20 subreddits with the most labels the flair text and css class were inspected. Flairs from a subreddit were accepted as an indicator of male or female gender if the text or css class specifically stated a variant of the query[10] and if the subreddit was deemed plausible to contain gender specific flairs. The subreddit clevelandcavs[11], for example, had a high amount of flairs with 'f', but inspection of the subreddit showed no plausible reason for this particular sports-team subreddit flair to be an indicator of gender, whereas for a subreddit such as AskWomen[12], a 'f' flair would be more plausible. Entries of subreddits of an erotic nature were also removed, as users of those subreddits often use 'throwaway' accounts, which is a Reddit specific term for accounts which are solely used to post and engage in specific threads or erotic subreddits. Even though some of these users get their gender verified by moderators of specific subreddits, they might not post anywhere else apart from certain specific subreddits and thus introduce extra noise due to very topic specific language if they would remain in the dataset. As a final step in the author filtering, bots (for example "auto moderator" which is a scripted moderator of a subreddit) and moderators of subreddits were removed by querying comments against text such as "*your submission has been removed*". An overview of subreddits which remained after filtering is given in table 1. Note that in table 1, asktransgender is in italics, it was included in the table to show that the current collection strategy can pick up gender signals from non-binary genders. For the analyses, however, *asktransgender* was not included.

For the non-SARC, there were extra steps. For each author that remained after the filters as described above, comments were gathered. It was ensured that the comments that were gathered did not appear in the subreddit where the author's flair was obtained from. This to ensure that a broader spectrum of information could be in the comments, as comments from the flaired subreddits have the potential to be quite biased in language use due to high frequency of specific topics discussed.

10 M/male, F/female, M/m and F/f
11 https://www.reddit.com/r/clevelandcavs/
12 https://www.reddit.com/r/AskWomen/

Table 1
Overview of subreddits which remained after filtering, sorted by number of flairs, descending.

SARC-gender	Non-SARC
AskMen	AskMen
AskWomen	AskWomen
AskMenOver30	AskMenOver30
AskWomenOver30	AskWomenOver30
datingoverthirty	datingoverthirty
sexover30	sexover30
askwomenadvice	askwomenadvice
ForeverAlone	GWABackstage
	asktransgender
	amiugly
	ForeverAlone
	DeadBedrooms
	40something
	RelationshipsOver35
	OkCupid

Table 2
Descriptives of the SARC-gender, non-SARC-100, non-SARC-200.

	SARC-gender	non-SARC-100	non-SARC-200
Male	2527 (77.3%)	5618 (63.7%)	3830 (66.13%)
Female	742 (22.7%)	3204 (36.3 %)	1962 (33.87%)
Total	3269 (100%)	8822 (100%)	5792 (100%)

To aid future research using Reddit in deciding the amount of comments needed for classification either due to time constraints or other reasons, we gather 100 and 200 comments per user, and treat these as different training datasets, called non-SARC-100 and non-SARC-200, and evaluate models trained on these different datasets. Lastly, to investigate the effectiveness of evaluating on either stand-alone comments or comments grouped by author, we employ two different versions of the SARC-gender, namely: the SARC-gendercom and the SARC-genderbatch. The former of which consists of single comments, the latter of which has comments grouped by authors.

3.4 Pre-Processing

To ensure clean data for modeling, several pre-processing steps were taken using regular expressions with the regular expression package in Python 3. Reddit uses Markdown[13] which is a specific type of text editing language. As this is not common punctuation people use, all these Markdown symbols were removed, preserving text

13 https://daringfireball.net/projects/markdown/

Table 3
Frequency of gender and sarcastic comments in the SARC-gender after pre-processing.

	Male	Female	Total
Authors	2527 (77.3%)	742 (22.7%)	3269 (100%)
Sarcastic comments	6988 (81.45%)	1592 (18.55%)	8580 (100%)
Non-sarcastic comments	7043 (81.53%)	1596 (18.47%)	8639 (100%)
Total for comments	14031 (81.49%)	3188 (18.51%)	17219 (100%)

within link (URL) contexts. Furthermore, newline and tab symbols were removed, as well as forward slashes (/). The raw Reddit comments can contain emoticons in unicode text, as such these were removed as well. Reddit specific text characteristics such as username mentions (denoted by /u/) or subreddit mentions (denoted by /r/) were removed, along with the user and subreddit names. Incorrect references to usernames and subreddit names were also removed. Lastly, applying the same text pre-processing pipeline to the comments of authors in the SARC-gender resulted in 12 comments being deleted. Inspection showed that these were non-sarcastic comments existing only of a subreddit or username mention. The final distribution of comments per gender can be seen in table 3.

3.5 Experimental procedure

All analyses in the following sections are implemented using Python 3.6.5 using Anaconda [14]. Packages used were Pandas (McKinney et al. 2010), Numpy (Oliphant 2006), Scipy (Oliphant 2007; Millman and Aivazis 2011) and Scikit-learn (Pedregosa et al. 2011). FastText[15] was implemented using a Scikit-learn wrapper.

There are two main parts: the gender information pre-test and the sarcasm prediction task. The pre-test is conducted to obtain an indication of how much gender information there is in the comments in the SARC. This indication is obtained by predicting gender from the SARC-gender and the best performing model will subsequently be used to predict gender labels on the remainder of the SARC. These predicted labels will then be encoded in the SARC comments and a new model will be trained to predict sarcasm.

3.6 Models

The models are obtained by using a continuous bag of words fastText model, which has proven to be a good baseline for text and sentiment classification (Joulin et al. 2016), and use character n-grams and word n-grams. Character and word n-grams have been used extensively in both sarcasm detection and author profiling to great results (Gonzalez-Ibanez, Muresan, and Wacholder 2011; Ferreira and Neto 2017; Sanchez-Perez et al. 2017).

14 https://anaconda.org/
15 https://fasttext.cc

FastText is a relatively new and fast text classifier and performs on par with several deep learning based classifiers. Input for this classifier is allowed to be raw text with corresponding labels. However, according to the fastText documentation, the tokenization process of fastText is relatively simple, therefore pre-processing and or more advanced tokenizers[16] are recommended. In our case, this meant using the SpaCy tokenizer (Honnibal and Johnson 2015) along with converting text to lower case and other steps as mentioned earlier. Afterwards, fastText processes the input in a hidden embedding layer and assigns class probabilities using a softmax function. When training on large amounts data, hierarchical softmax is recommended for efficiency when classes are balanced, but at the possible cost of accuracy loss[17] when the classes are unbalanced. Since the test and training data for the pre-test do not have balanced classes, we opt to use the normal softmax function. Lastly, due to the nature of the stochastic gradient descent (SGD) algorithm variation called hogwild (Recht et al. 2011), used by fastText, some randomness will exist in the results and this can not be controlled by setting a specific random state.

All fastText models are trained and tuned using 5-fold Cross Validation. To find the best hyper-parameters, random search (Bergstra and Bengio 2012) was performed (due to time constraints a full exhaustive gridsearch was unable to be performed). Hyper-parameters that were tuned are: learning-rate (between 0.001 and 1), dimensions (dim) between 10 and 300, size of the context window (ws) between 1 and 3, epochs (between 1 and 50), word and character n-grams between 1 and 6. The following parameters were left at default: bucket size (2M), learning update rate (100), minimum word count (5), number of negatives at default (5). Number of threads were set at 6.

3.7 Pre-test: Gender information in the SARC

To answer the research question "Is there gender information in sarcastic comments" a binary classification experiment is conducted. Gender is treated as binary in the current study, with male (0) and female (1). The baseline for this experiment is the majority baseline, as men are more prominent in the test-set the majority baseline is 77.3%. The idea is that if the models perform better than majority baseline, then there is evidence that there is gender information in sarcastic comments. A model with good performance will lead to more reliable prediction of gender labels on the remainder of the SARC, which in turn will influence the performance obtained in the sarcasm classification task.

As seen in table 2 and evidenced by the majority baseline, there is a skew in the train and test-sets, however for all of these datasets the difference in skew in gender is not large between the training sets and test set and generally follows the same direction, with more men having self-reported their gender than women. However, in terms of comments in the SARC-gender (test-set), there is a larger skew with men having more sarcastic comments (and more comments overall) than women. Many studies in the field of author profiling and sarcasm detection use accuracy as the evaluating metric due to its ease of interpretation. However, as our data shows a skew, the accuracy metric can not be used to reliably interpret the result. The skew-ratio for our dataset is 4.4 in terms of comments and 3.4 in terms of gender which is relatively close to the normal functioning of the F1-score (Jeni, Cohn, and De La Torre 2013). However, the F1-score is influenced by unbalanced classes. Since the current experiment does not have a

16 https://github.com/facebookresearch/fastText/blob/master/python/README.md
17 https://fasttext.cc/docs/en/faqs.html

Table 4
Frequency of gender and comments after gender label prediction on the remainder of the SARC.

	Male	Female	Total
Authors	179034 (70.68%)	74256 (29.32%)	253290 (100%)
Sarcastic comments	383942 (77.28%)	112843 (22.72%)	496785 (100%)
Non-sarcastic comments	384250 (77.35%)	112507 (22.65%)	496757 (100%)
Total for comments	768230 (77.32%)	225363 (22.68%)	993593 (100%)

preferential class to predict, we use the macro-averaged F1-score, since it puts an equal weight on both classes which allows for a fairer inspection of the importance of the minority class (Schütze, Manning, and Raghavan 2008).

The structure of the pre-test is as follows: first we compare the performance of models trained on the non-SARC-100 and 200 and evaluate these on the SARC-gendercoms in three conditions: non-sarcastic, sarcastic, and mixed messages. To further illustrate the differences between the non-SARC-100 and 200 in terms of gender prediction performance, we include the micro F1-score as well. The best performing model will be evaluated on the SARC-genderbatch in the three conditions. Lastly, POS-tags will be constructed and supplied to a model without character n-grams to see if it improves the gender classification. POS-tags have been shown to improve gender classification in the work of Soler Company (2016) where they performed better than models using only character or word n-grams. An overview of the best performing models with hyperparameters and features is shown in table 5.

3.8 Sarcasm classification task

To answer the second research question "can gender information improve sarcasm classification" a binary classification experiment is conducted on the SARC, with sarcastic comments being labeled as 1, and non-sarcastic as 0. The structure of this experiment is as follows: we use the best-performing gender prediction model to predict gender-labels for the remainder of the users in the SARC. The best performing model was the POS-tagged model, therefore POS-tags are constructed for the entire SARC after the same text pre-processing steps as mentioned earlier. Once again we group comments by authors, as our best model had the highest performance in the batched setting. In doing so it is ensured that single comments by authors are not predicted as a different gender by our model. Afterwards, their gender was predicted. The result of the gender prediction is visible in table 4.

Once the labels were predicted, we encode gender explicitly at the beginning of the comments. Gender is encoded in comments with POS-tags as well as comments without POS-tags, and we treat these as two different datasets. We use two models which are trained and tested with gender encoding. The first is a model without POS-tags in the comments and this uses only character and word n-grams, both of these features have been used in sarcasm detection before (Gonzalez-Ibanez, Muresan, and Wacholder 2011). The second is a model with POS-tags in the comments and does not use character n-grams, but does use word n-grams. Since our best performing gender

Table 5
Optimal model hyperparameters of best performing models in the gender pre-test.

Test set:	SARC-genderbatch	SARC-genderbatch
Training data:	Non-SARC-200	Non-SARC-200
Condition:	Non-sarcastic only	Mixed
Parameters		
Dim	100	300
Epoch	5	5
Lr	0.1	0.1
Ws	5	5
minN	1	0
maxN	3	0
Word n-gram	(3,5)	(3,5)
POS-tags	No	Yes

prediction model used POS-tags and these tags have been used in sarcasm detection (Bouazizi and Ohtsuki 2016) we include this model as well.

We encode gender in a fashion similar to Johnson et al. (2016), where they explicitly encoded the language of a comment in order to help with machine translation of languages, to great effect. The reasoning behind this is that when we explicitly encode gender information, the model will learn to associate gender with certain characteristics in the comments and could therefore perform better on the sarcasm classification task. As an example, a comment such as "It sure looks like this is going to work" can become "<male> It sure looks like this is going to work". Afterwards, the comments were removed from the author grouping and split 80:20 into a train and test set, and the analysis procedure described under subsection 3.6 continued. The models were evaluated with macro F1 score (to remain consistent with the pre-test) and accuracy (as the SARC baselines are evaluated on accuracy). Lastly, the gender encoded models were compared to a fastText model that was trained and tested without explicitly encoded gender information and the best performing simple linear classifiers presented in Khodak, Saunshi, and Vodrahalli (2018), as well as the majority baseline of 50%.

4. Results

4.1 Pre-test: Gender information in the SARC

To obtain insight if there are gender differences in sarcasm and how much gender information there is in the users that appear in the SARC, a binary classification experiment was conducted to predict gender from non-sarcastic, sarcastic and a mix of comments, trained on 100 or 200 comments batches and evaluated on the SARC-gendercom and SARC-genderbatch on three conditions; mixed messages, sarcastic only and non-sarcastic only. The tables for this subsection are tables 5, 6 and 7. Table 5 shows the hyperparameters corresponding to the best performing gender pre-test models, as well as their train and test data and the condition and features used. Table 6 shows the best gender classification models, their features and the condition wherein the best performance was achieved, as well as which test-data they were tested on. In table 6 the

micro-F1 score is included as well, to aid in illustrating the difference in performance when training on 100 or 200 comments per author. Table 7 shows performance on individual classes of the best performing models, trained on either the non-SARC-100 or 200, and the condition on which they achieved the best scores. For all the experiments in this section, the baseline is the majority baseline of 0.773.

4.1.1 Non-SARC-100 versus non-SARC-200 on SARC-gendercoms. To determine the best strategy in terms of training data on gender prediction performance, we first evaluated the effectiveness of training on 100 and 200 comment batches and evaluated on the SARC-gendercoms. The models for this test used only character and word n-grams. As is shown in table 6, the best results were found in the non-sarcastic messages condition. However, the results were overwhelmingly in favour of using the non-SARC-200 to train, as the non-SARC-100 performed poorly having a macro-F1 score of 0.46, versus the non-SARC-200's macro-F1 of 0.53.

Furthermore, the non-SARC-100 was unable to properly detect female signals with an F1 score of 0.27, as visible in table 7. From this table it can be inferred that even though the non-SARC-100 achieves high individual scores for male (0.90), it is heavily influenced by having more labels of the majority class (illustrated by high discrepancy in macro/micro-F1 in table 6, 0.46 versus 0.81, respectively). This difference in performance can also mean that there is insufficient differentiating gender information in the non-SARC-100, since the non-SARC-200 gathers more female signals. Illustrated by female F1 scores of 0.31 versus the non-SARC-100 score of 0.27. That the models trained on 200 messages per user perform better in author profiling tasks is also found in the work of Volkova et al. (2015), which is in line with our expectations. However, both models still performed below majority baseline. As the non-SARC-100 model performed poorly, this training data was further excluded from training models.

4.1.2 Non-SARC-200 on SARC-gendercoms and SARC-genderbatch. To determine the best strategy in terms of test-data on gender prediction performance, the performance of a model trained on the non-SARC-200 was evaluated on SARC-gendercoms and the SARC-genderbatch. The metrics show that grouping the comments by author in the SARC results in a better performance, increasing the macro F1-score to 0.58, as opposed to the SARC-gendercoms macro F1-score of 0.53. These scores are visible in table 6. For each of the other conditions the difference in the F1-Macro only differed by +- 0.1, with the mixed messages condition always scoring 0.1 lower than non-sarcastic messages only, and sarcastic messages only always scoring 0.1 lower than mixed messages. These results indicate that the best strategy for gender prediction is to evaluate on the SARC-genderbatch. However, the models evaluated on the SARC-genderbatch still performed below majority baseline. To test if POS-tags can improve gender prediction, these will be constructed on the non-SARC-200 and evaluated on the SARC-genderbatch.

4.1.3 POS-tagged on the SARC-genderbatch. As seen in table 6, the POS-tag model performed best in the mixed condition, improving the macro F1 score by 0.01, compared to the best performing non-SARC-200 model. Other conditions had a lower macro-F1 score, with sarcastic only at 0.54, and non-sarcastic at 0.56. In terms of individual male and female classification performance we see that the POS-tags improve both male and female F1-scores. As visible in table 7; compared to only character and word n-grams, the male F1 score increases by 0.02, and the female F1 score increases by 0.09. That the POS-tagged model performed better than character and word n-grams is in line with our expectations and the work of Soler Company (2016). However, the models

Table 6
Best gender classification models, their training data and test-data, condition and features.

Model	SARC-gender version	Macro F1	Micro-F1
Training data: Non-SARC-200	Comments	0.53	0.63
Features: Word, character n-grams	Batches	0.58	0.68
Condition: Non-sarcastic only			
Training data: Non-SARC-200	Comments	*Not tested*	
Features: POS-tags, word n-grams	Batches	**0.59**	0.67
Condition: Mixed comments			
Training data: Non-SARC-100	Comments	0.46	**0.81**
Features: Word, character n-grams	Batches	*Not tested*	
Condition: Non-sarcastic only			
	Majority	0.773	

Table 7
Best performing Non-SARC-100 versus Non-SARC-200 in male and female classification. The POS-tag model is evaluated on SARC-genderbatch, the rest on SARC-gendercoms.

Model	Gender	F1
Non-SARC-100	Male	**0.90**
non-sarcastic messages	Female	0.27
Non-SARC-200	Male	0.75
non-sarcastic messages	Female	0.31
Non-SARC-200	Male	0.77
POS-tags, mixed	Female	**0.40**
	Majority	0.773

still perform below majority baseline. Lastly, the mixed condition was considered the most important for the current study, as sarcasm detection in practice will also feature non-sarcastic messages. Therefore, this model was used to predict gender labels on the remainder of the SARC.

4.2 Sarcasm classification task

To investigate if gender information can enhance sarcasm detection, gender labels were predicted on the remainder of the SARC and afterwards gender was explicitly encoded in the comments. Thereafter a a binary classification experiment was conducted on the balanced SARC with gender encoding. To compare the gender encoded models to a normal fastText model, a normal fastText model was also trained and tested. Table 8 shows the optimal hyperparameters for the best performing models. Table 9 shows the classification results, as well as the linear classifier baselines for the balanced SARC from Khodak, Saunshi, and Vodrahalli (2018). As they used accuracy, we incorporate

Table 8
Optimal model hyperparameters for the sarcasm classification task.

	Gender encoding	No gender encoding	Gender + POS-tags
Dim	300	300	300
Ws	3	1	1
Epoch	50	5	5
Lr	0.05	0.01	0.01
MinN	1	2	0
MaxN	5	5	0
Word n-gram	(2,3)	(3,6)	(3,6)

Table 9
Best performing sarcasm classification models. SARC-baselines are the linear classifier baselines reported in Khodak, Saunshi, and Vodrahalli (2018).

Model	Accuracy	Macro-F1
Gender encoded	0.705	0.7
Gender encoded + POS-tags	0.700	0.7
No gender encoding	0.709	0.7
SARC-baselines		
Bag-of-Words	0.732	
Bag-of-Bigrams	0.758	
Sentence embedding	0.71	
Baseline	0.5	

accuracy in table 9 as well besides the macro F1-score, to stay consistent with metrics used in the pre-test.

From table 9 it is visible that the gender encoding with and without POS-tags, and no gender encoding trained models perform above baseline, with an accuracy of 0.705, 0.700 and 0.709 respectively. However, these do not perform any better than the baselines reported by Khodak, Saunshi, and Vodrahalli (2018). The models used in the current study do not differ substantially in individual class scores (sarcastic or non-sarcastic comments) and these individual class scores are therefore not included in table 9. These results are not unexpected as the pre-test demonstrated that there was not a lot of gender information in the profiles on the SARC. Predicting gender labels on the remainder of the SARC therefore would not yield the most reliable gender labels.

5. Discussion

The objective of this research was to analyze if there is gender information in sarcasm, and subsequently if this can enhance sarcasm detection on Reddit. This was approached in two phases. First it was investigated if machine learning models can detect gender information in sarcastic messages, this was tested by predicting gender from a set of sar-

castic comments, non-sarcastic comments and a combination thereof. Features used in this phase were character and word n-grams and Part-Of-Speech tags. Secondly, it was investigated if explicitly encoding gender information can enhance sarcasm detection. To this end, the best performing model from the first phase was used to predict gender labels for the remainder of the users, after which these were explicitly encoded into the comments. In the second phase there were three fastText models: a word and character n-gram model, a word n-gram and POS-tag model, and a word and character n-gram model without gender encoding. All three models were trained and tested in a binary classification experiment.

To answer the first research question, we look at the results of the gender pre-test. It was expected to find sufficient gender information in the comments to efficiently classify gender on the SARC by using word and character n-grams as well as POS-tags, however the results did not corroborate this. Even though word and character n-grams and combinations with other features have been used in earlier work to predict gender (Ferreira and Neto 2017; Sanchez-Perez et al. 2017), these were, in combination with the current study set-up and models used, unable to effectively detect possible gender signals. This could mean that there is simply not enough gender information in the current profiles and comments on the SARC. Higher amount of word n-grams were a common characteristic of the best performing models, as is the case in (Sanchez-Perez et al. 2017). The expectation that POS-tags would improve gender prediction was in line with the results and the work of Soler Company (2016). Since the POS-tag model performed the best in the mixed condition, this was especially important, as in practice sarcastic comments can appear next to non-sarcastic ones. Moreover, the POS-tag model resulted in the highest improvement of female F1-score compared to the other models. However, the POS-tag model, as the other models, did not improve the classification sufficiently for it to perform above majority baseline. Lastly, the best performing models were trained on 200 messages per user and tested on comments grouped by author in the SARC.

During the gender pre-test, several subtasks were performed. These were: comparing the performance of models trained on 100 and 200 messages, and comparing evaluating on single comments as well as batched comments of the SARC. Since Reddit has not been used extensively in research, these subtasks were deemed important to aid future research on this platform in deciding their train and test data structure. The first subtask, comparing 100 and 200 messages per author showed that models trained on the batches of 100 comments per user performed extremely poor, indicating that, even though there are more labeled users, the gender information they contain is quite low. Other research, such as the work of Volkova et al. (2015) using Twitter, also leveraged 200 messages per user to great effect. This suggests that gathering 100 messages per user is insufficient to capture author-specific traits, and 200 messages per author allows for more encoding of author characteristics, even on Reddit. The second subtask, comparing evaluating on single comment or comments grouped by author showed that the best results were obtained using the comments grouped by author. Since the SARC contains sarcastic messages, single messages could influence gender prediction. By leveraging more comments by an author and treating those as a single batch, one could encode more gender and author specific information which in turn could capture specific characteristics of an author, even in sarcasm-tinted language. This is in line with the successful work of Amir et al. (2016) in sarcasm detection, wherein, among other features, user-embeddings were used to predict sarcasm.

To answer the second research question, we examine the results of the sarcasm classification task. It was expected that gender information could enhance sarcasm

classification. Three fastText models were compared in their effectiveness on predicting sarcasm in this binary experiment. Two models were trained and tested on the explicit gender encoded messages in the SARC, one with only character and word n-grams and the other with POS-tags and word n-grams. The last model was trained and tested without gender encoding and included character and word n-grams. All three models performed above majority baseline in terms of accuracy and F1-score. The gender encoded models did not outperform the non-gender encoded model and the linear classifier baselines reported in the work of Khodak, Saunshi, and Vodrahalli (2018). It is not surprising that explicitly encoding gender has little effect in the current study, as the gender information pre-test proved that predicting gender for the SARC users is difficult. Since we used the best performing model to predict the gender of the remaining authors in the SARC the resulting labels are unreliable, which in turn would influence the models learning ability on the sarcasm classification task. This could also be a reason for the gender encoded models to perform slightly worse than without gender encoding, meaning that the inclusion of gender in comments added noise. However, as these differences are minuscule, another explanation could be that these differences can be attributed to randomness when using fastText or randomized search (Bergstra and Bengio 2012). As expected, we did see that word and character n-grams in combination with fastText word-embeddings are effective in predicting sarcasm in this experiment, as all models still performed well above majority baseline. This is in line with work by Gonzalez-Ibanez, Muresan, and Wacholder (2011) who used n-gram variations, Ghosh, Guo, and Muresan (2015) who used word-embeddings and Bouazizi and Ohtsuki (2016) who used POS-tags, to predict sarcasm. Furthermore, this once more reinforces the notion that fastText serves as an efficient baseline classifier in sentiment analysis as reported in Joulin et al. (2016), as it performs, in one of the most difficult tasks of sentiment analysis; sarcasm detection, on par with the linear baseline classifiers reported in Khodak, Saunshi, and Vodrahalli (2018).

To summarize, the answer to both research questions is that predicting gender from sarcasm is difficult, and that it does not necessarily enhance sarcasm detection, at least in the current study set-up. In terms of data analysis and collection strategies on Reddit, we have shown that using 200 messages per author results in better performance on the gender prediction task and that evaluating on SARC-author profiles (messages grouped by author) results in better performance. These results can help researchers decide their approach for conducting more research on Reddit or on the SARC.

5.1 General observations

An interesting aspect of the data used in the current study is that, with usage of only a few months of Reddit comments, we were able to obtain a relatively large number of gender labels for authors in the SARC, compared to the years of comments with which the SARC was constructed. In terms of gender distribution, both the non-SARC authors and SARC-authors followed the same trend, with more men having self-reported their gender on Reddit. Moreover, in terms of comments in the SARC, the amount of comments by male authors far exceeded the amount of female comments for both sarcastic and non sarcastic comments. Research by Rockwell and Theriot (2001) showed that, based on self-assessments, men were more sarcastic overall. The current distribution of the SARC gender seems to support this, with 6998 sarcastic comments by men and 1592 by women. However, this should be viewed in a nuanced way as the current iteration of the SARC-gender uses only a few months of comments to acquire labels. Lastly, we see that the distribution of the predicted gender labels on the rest of the SARC

remains approximately equal to the SARC-gender labels. If our models were capable of properly detecting female signals or trained on more female text information (such as in a balanced setting), maybe this distribution would end up differently.

5.2 Future work

Future work could focus on using gender prediction trained word-embeddings to predict sarcasm, as word-embeddings have proven useful in sarcasm detection (Ghosh, Guo, and Muresan 2015), perhaps these can be used to train a sarcasm detection model as well. Due to the relative effectiveness of the data collection strategy one could use a equivalent strategy to obtain different demographic information such as age by only querying flairs. One could also opt to extend this strategy and apply the simple queries method by Emmery, Chrupała, and Daelemans (2017) for salient information in the body of comments to obtain labels. This could allow researchers to investigate Reddit and its users in a more in-depth manner, or use different demographic information to supplement sarcasm research on the SARC. Moreover, in terms of research on the SARC, it could be interesting to investigate subreddit language use in relation to sarcasm, or quantify a subreddit's tendency to be sarcastic to predict sarcasm, since Khodak, Saunshi, and Vodrahalli (2018) found that the politics subreddit had a high amount of sarcastic comments. Lastly, since sarcasm is a rare occurrence, more would could be done on predicting sarcasm from unbalanced datasets.

5.3 Limitations

The current study has a skewed dataset in terms of gender labels. This could be due to only using a few months of Reddit comments to obtain labels. This resulted in few female labels for the SARC as well as in the training data, perhaps with more labels better predictions could be made. Perhaps a balanced setting for training data would improve classification performance. However, this would imply that one expects a priori that the representation of gender on Reddit is balanced. Since, at least in terms of self-reports of gender, there is no indication that this is the case, the current study decided not to balance the train or test data. It could be the case that men on Reddit simply self-report their gender more often than women in their user flairs. Perhaps supplementing flair-labels with simple queries (Emmery, Chrupała, and Daelemans 2017) could increase the female label frequency. Furthermore, even though there was an unbalanced dataset, there an extensive effort was made to ensure that the labels that were collected were as accurate as possible. In terms of pre-processing it was not ensured that the language in the training data was limited to English. Reddit can contain comments in different languages, however we expect these comments to be in the minority since English is the primary and official language used on Reddit. Lastly, the current work did not evaluate on the unbalanced test-set of the SARC, since sarcasm is a rare occurrence more work could be done investigating the results on the unbalanced set. We do expect the current best performing models to suffer greatly in performance, as did the performance of the linear classifiers in the work of Khodak, Saunshi, and Vodrahalli (2018) on an unbalanced test set.

6. Conclusion

The current study focused on using gender information to enhance sarcasm detection on the Self-Annotated Reddit Corpus. Gender labels were obtained by deploying a

collection strategy inspired by Emmery, Chrupała, and Daelemans (2017) on raw Reddit comments by focusing on user flairs. Extensive label inspection was performed to ensure quality labels before being added to authors in the SARC. Afterwards it was investigated whether a combination of word and character n-grams as well as POS-tags can predict gender from sarcastic comments using fastText models. Gender prediction proved a difficult task, with the best performing model having a combination of word n-grams and POS-tags achieving a macro F1 score of 0.59, which remained below majority baseline (0.773). Subsequently, this best model was used to predict gender labels on the SARC. These labels were encoded explicitly into the comments following the strategy by Johnson et al. (2016). Sarcasm prediction from explicit gender encoded messages did not perform better in terms of accuracy and macro F1-score than a model based on messages without explicit gender encoding. The gender encoded model with character and word n-grams had an accuracy of 0.705 and macro F1-score of 0.7 and the gender encoded model with POS-tags reached an accuracy of 0.7 and macro F1-score of 0.7. The model without gender encoding had an accuracy of 0.709 and macro F1-score of 0.7. Furthermore, these models did not perform better than the linear baseline classifiers provided by Khodak, Saunshi, and Vodrahalli (2018), but did perform above majority baseline (0.5). Therefore, in the current study set-up, gender was unable to be effectively predicted from sarcastic messages and gender information did not enhance sarcasm detection.

Since Reddit has not been studied extensively, we provided insights in different strategies to obtain the best possible results. To this end, we analyzed the effects of training on 100 versus 200 comments per author, and found that 200 comments per author offers the best performance in terms of gender performance. Furthermore, we provided a detailed data collection strategy and steps to ensure good quality self-reported labels using the current collection strategy. Moreover, our collection strategy can pick up on non-binary gender, but these were excluded from the analyses. Lastly, a comparison has been made between testing on single comments on the SARC versus comments grouped by author in terms of gender prediction. The best strategy in the current research was to evaluate on comments grouped by authors. This can be useful to evaluate other demographic information that might be in the SARC.

References

Abbasi, Ahmed, Hsinchun Chen, and Arab Salem. 12 sentiment analysis in multiple languages: Feature selection for opinion classification in web forums.

Amir, Silvio, Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976*.

Argamon, Shlomo, Moshe Koppel, James Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52:119–123.

Bamman, David and Noah A. Smith. 2015. Contextualized sarcasm detection on twitter. In *ICWSM*.

Bergstra, James and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Bouazizi, M. and T. Otsuki Ohtsuki. 2016. A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488.

Carvalho, Paula, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 53–56, ACM, New York, NY, USA.

Culotta, Aron, Nirmal Kumar Ravi, and Jennifer Cutler. 2016. Predicting twitter user demographics using distant supervision from website traffic data. *Journal of Artificial Intelligence Research*, 55:389–408.

Davidov, Dmitry, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Association for Computational Linguistics, Stroudsburg, PA, USA.

Dress, Megan L, Roger J Kreuz, Kristen E Link, and Gina M Caucci. 2008. Regional variation in the use of sarcasm. *Journal of Language and Social Psychology*, 27(1):71–85.

Drucker, Ari, Ofer Fein, Dafna Bergerbest, and Rachel Giora. 2014. On sarcasm, social awareness, and gender. *Humor*, 27(4):551–573.

Emmery, Chris, Grzegorz Chrupała, and Walter Daelemans. 2017. Simple queries as distant labels for predicting gender on twitter. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 50–55.

Estival, Dominique, Tanja Gaustad, Son Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for english emails.

Ferreira, Rosalvo and Rosalvo Neto. 2017. Using character n-grams and style features for gender and language variety classification.

Ghosh, Debanjan, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1003–1012, Association for Computational Linguistics.

Go, Alec, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12).

Gonzalez-Ibanez, Roberto, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. pages 581–586.

Hazarika, Devamanyu, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. *arXiv preprint arXiv:1805.06413*.

Honnibal, Matthew and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Association for Computational Linguistics, Lisbon, Portugal.

Jeni, László A, Jeffrey F Cohn, and Fernando De La Torre. 2013. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 245–251, IEEE.

Johnson, Melvin, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Joshi, Aditya, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5):73:1–73:22.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Khodak, Mikhail, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Linguistic Resource and Evaluation Conference (LREC)*.

Kolchinski, Y Alex and Christopher Potts. 2018. Representing social media users for sarcasm detection. *arXiv preprint arXiv:1808.08470*.

Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Kreuz, Roger J. and Sam Glucksberg. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, 118(4):374–386.

L. Newman, Matthew, Carla J. Groom, Lori D. Handelman, and James Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes - DISCOURSE PROCESS*, 45:211–236.

Lakoff, Robin. 1973. Language and woman's place. *Language in Society*, 2(1):45–80.

Li, Jiwei, Alan Ritter, and Eduard Hovy. 2014. Weakly supervised user profile extraction from twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 165–174.

Liebrecht, CC, FA Kunneman, and APJ van Den Bosch. 2013. The perfect solution for detecting sarcasm in tweets# not.

Maynard, D.G. and M.A. Greenwood. 2014a. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. The LREC 2014 Proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Maynard, Diana and Mark A. Greenwood. 2014b. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *LREC*.

McKinney, Wes et al. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, Austin, TX.

Mensikova, Anastasija and Chris A Mattmann. 2018. Ensemble sentiment analysis to identify human trafficking in web data.

Millman, K Jarrod and Michael Aivazis. 2011. Python for scientists and engineers. *Computing in Science & Engineering*, 13(2):9–12.

Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011, Association for Computational Linguistics.

Mosteller, Frederick and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Neal, Tempestt, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Comput. Surv.*, 50(6):86:1–86:36.

Oliphant, Travis E. 2006. *A guide to NumPy*, volume 1. Trelgol Publishing USA.

Oliphant, Travis E. 2007. Python for scientific computing. *Computing in Science & Engineering*, 9(3).

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. *EMNLP*, 10.

Park, Gregory, David Bryce Yaden, H Andrew Schwartz, Margaret L Kern, Johannes C Eichstaedt, Michael Kosinski, David Stillwell, Lyle H Ungar, and Martin EP Seligman. 2016. Women are warmer but no less assertive than men: Gender and language on facebook. *PloS one*, 11(5):e0155885.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Pennebaker, James W., Matthias R. Mehl, and Kate Niederhoffer. 2003. Psychological aspects of natural language. use: our words, our selves. *Annual review of psychology*, 54:547–77.

Raghav, S. and E. Kumar. 2017. Review of automatic sarcasm detection. In *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, pages 1–6.

Rangel, Francisco, Paolo Rosso, Manuel Montes-y Gómez, Martin Potthast, and Benno Stein. 2018. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working Notes Papers of the CLEF*.

Rangel, Francisco, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working Notes Papers of the CLEF*.

Read, Jonathon. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Association for Computational Linguistics, Stroudsburg, PA, USA.

Recht, Benjamin, Christopher Re, Stephen Wright, and Feng Niu. 2011. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems*, pages 693–701.

Reyes, Antonio, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Rockwell, Patricia and Evelyn M Theriot. 2001. Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis. *Communication Research Reports*, 18(1):44–52.

Sanchez-Perez, Miguel, Ilia Markov, Helena Gomez Adorno, and Grigori Sidorov. 2017. Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus (preprint version). pages 145–151.

Sarawgi, Ruchita, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86, Association for Computational Linguistics.

Sboev, Alexander, Tatiana Litvinova, Dmitry Gudovskikh, Roman Rybka, and Ivan Moloshnikov. 2016. Machine learning models of text categorization by author gender using topic-independent features. *Procedia Computer Science*, 101:135–142.

Schler, Jonathan, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Schütze, Hinrich, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press.

Schwartz, H Andrew, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Simaki, Vasiliki, I Mporas, and Vasileios Megalooikonomou. 2016. Evaluation and sociolinguistic analysis of text features for gender and age identification. 9:868–876.

Soler Company, Juan. 2016. Use of discourse and syntactic features for gender identification.

Tepperman, Joseph, David Traum, and Shrikanth Narayanan. 2006. Yeah right: Sarcasm recognition for spoken dialogue systems.

van Dalen, Reinard, Léon Melein, and Barbara Plank. 2017. Profiling dutch authors on twitter: Discovering political preference and income level. *Computational Linguistics in the Netherlands Journal*, 7:79–92.

Verhoeven, Ben, Walter Daelemans, and Barbara Plank. 2016. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)/Calzolari, Nicoletta [edit.]; et al.*, pages 1–6.

Volkova, Svitlana, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *AAAI*, pages 4296–4297.

Wallace, Byron C., Do Kook Choe, and Eugene Charniak. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, volume 1, pages 1035–1044, Association for Computational Linguistics (ACL).

Wang, Zelin, Zhijian Wu, Ruimin Wang, and Yafeng Ren. 2015. Twitter sarcasm detection exploiting a context-based model. In *International Conference on Web Information Systems Engineering*, pages 77–91, Springer.

Ruben Leonard van de Kerkhof Gender information absolutely enhances sarcasm detection (/s)

Zheng, Yunpei, Lin Li, Luo Zhong, Jianwei Zhang, and Jinhang Liu. 2018. Using sentiment representation learning to enhance gender classification for user profiling.