Mining of Misogyny Categories in Tweets

s1038931
Radboud University, Nijmegen, Netherlands

ABSTRACT

Misogyny, which is a common phenomenon in social networks, is detrimental to women in many ways. It may discourage them from using online communities and sometimes takes a form of life-threatening. However, only recently an Automatic Misogyny Identification (AMI) taks specifically devoted to misogyny identification and categorisation was organised. In this paper, I examine misogyny categories present in the data set of this task and try to discover new categories by building the most representative features set and clustering misogynous tweets. In addition to the categories defined by AMI organisers, two more categories were discovered. Also, more insight into the difficulty of applying NLP methods to research in hate speech was gained.

1 INTRODUCTION

Despite the progress made towards gender equality in recent centuries, prejudice and violence against women are still common and ubiquitous. And the Internet is no exception. Misogyny in online social media not only silences women but also may lead to very serious consequences. For instance, victims of Gamergate received a huge amount of rape and death threats and eventually, some of them were forced to leave their houses [10]. Moreover, it has been shown that beliefs in stereotypes about women correlate with acceptance of violence against them [4]. In social networks posts and trends spread at lightning speed, therefore without any control social media can contribute to the rise of misogyny level in the society.

Currently, online platforms acknowledge the harmful influence of hate speech and prohibit it. However, its monitoring and removal are still manual and inefficient. Automation is highly desirable to speed up the process and set people free from this unpleasant job.

The results of the resent studies devoted to different aspects of hate speech detection reveal that it is a highly non-trivial task. Moreover, [2] shows that the performance of the current state-of-the-art methods is substantially overestimated and the most popular data sets in the field tend to be biased. Therefore, to build reliable automatic tools more research and more insight into different forms of hate speech is needed.

So far misogyny received rather little attention from the research community. Only recently two data sets intended specifically for misogyny detection and categorisation were released [6], [7]. Misogyny detection turned out to be a feasible task, whereas identifying its categories was way more difficult. In this paper, I investigate the sources of this difficulty by clustering the misogynous tweets and checking whether it is possible to extract misogyny categories defined for AMI task in an unsupervised manner. In addition, I check if the data set contains more misogyny categories by setting a number of clusters higher that number of known categories and by clustering the largest defined category.

TxMM'19-'20, Radboud University, Nijmegen, Netherlands

2 RELATED WORK

2.1 Hate speech

In the past few years, there has been a surge of interest in automatic hate speech detection. Quite a few studies have been carried out and the researchers constructed diverse data sets. [9] built a data set for abusive language detection labelled with different subcategories (hate, derogatory language and profanity) and evaluated the effectiveness of a wide range of features. Davidson et al. work [5] shows that it is hard to reliably distinguish between hate speech and other types of offensive language on Twitter. Generally, their model tended to misclassify tweets with general profanity as hate speech. SemEval-2019 workshop presented a task to detect offensive language and offence target using a novel hierarchical three-level annotation scheme and a data set created specifically for this task [14]. The most recent and detailed survey about the current state of the field was performed by Fortuna et. al. [8].

2.2 Misogyny

Among all the studies related to hate speech and offensive language, only a few papers looked into misogyny. The first study which built a model to detect sexism in Twitter and analysed the discriminative impact of different features is [12]. SemEval-2019 Task 5 [3] was aimed at detection of hate speech against immigrants and women in Twitter, however, it only labels a tweet as hateful or non-hateful and does not annotate sexism and racism separately. Finally, [6] and [7] challenges were about misogyny detection exclusively. Both use the same annotation scheme, which includes misogyny categories and target labels. The goal of subtask B of these challenges was to determine a type of misogynistic behaviour and classify its target. Macro-average F1-score was used as an evaluation metric. It turned out to be an extremely difficult task: at AMI only two teams managed to beat an organizers' baseline for the English language data set. And the baseline score was relatively low: 0.370 for subtask B and 0.342 for categories identification.

2.3 Main challenges discovered

There are a few problems in hate speech detection research. First, there is no standard benchmark data set and as pointed out in [13] there is still no agreement about a unified annotation scheme. Even more serious problem is that a model trained on one data set generally performs badly on another data set [2]. The reason for that is overfitting and sampling issues: it turned out that in the most popular data set for hate speech detection [12] most of the tweets were generated by only a few users. This demonstrates the low quality of data sets, the bias of state-of-the-art models and that the reported performance is significantly overestimated. Hence, the proposed models cannot be applied in practice yet. Lastly, hate

speech detection has many known challenges, e.g often it implies irony detection, it could be context-depended and expressed in entangled ways. Therefore, further research is needed to build robust models.

As we can see, the most prominent work addressed automatic hate speech detection by formulating this problem as a supervised classification task and the reliable solution still has not been found. Thus, it is sensible to try unsupervised machine learning methods to get more insight into the semantic gap characteristic for this task and possibly collect some useful information. In particular, clustering of misogynous tweets may help to discover unobserved misogynous categories and show how feasible their automatic extraction is.

3 DATA

I chose to work with the English part of the data set created for AMI task from EVALITA 2018 [6] evaluation campaign of NLP. It contains 4001 tweets in the training set and 1001 tweets in the test set. Each tweet marked as misogynous has a category assigned. The organizers identified five misogyny categories: : **stereotype**, **dominance**, **derailing**, **sexual harassment**, **discredit**. The detailed annotation scheme can be found in [6].

The data set is available for free, however, it is distributed via private google group. To get the data, one needs to request to join the group and get approval from the administrators.

Since I want to examine misogyny categories, I merge test and training set and work only with tweets marked as misogynous. There are 2245 such tweets in the data set. It is important to note, that categories are imbalanced. The number and percentage of tweets in each category are given in Table 1.

Table 1: Distribution of misogyny categories. The second row shows the percentage of each category in the data set.

Discredit	Sexual harassment	Stereotype	Dominance	Derailing
1155	396	319	272	103
51.4%	17.6%	14.2%	12.1$	4.6%

4 APPROACH

First of all, I performed a quick and simple sanity check of the categories selected by the organizers by comparing the most frequent words in each category. Stop words and punctuation marks were removed but emojis and hashtags are left.

The second step is clustering of the tweets, evaluation of the clusters and analysis of features that prevent the tweets to fall into the "correct" clusters. For general pre-processing tweets were lowercased, links, usernames and hashtags were replaced with LINK, USERNAME and HASHTAG tokens respectively. Also, extra white spaces and retweet prefix (RT) were removed. Considering the methodology of [9] and [1], I investigate the following features:

N-grams: Tokens unigrams, bigrams and trigrams as well as characters n-gram ($n = 3, 4, 5$). Both tokens and characters n-grams were extracted without punctuation marks, which were added to linguistic features set instead. Tokens n-grams which purely consist of stop words were disregarded and also stop words were removed before character n-grams extraction. In addition, rare words (those which occur less than 3 times in tweets set vocabulary) were added to stop words set during tokens n-grams extraction. This was done because rare words do not contribute to measuring similarity. However, rare words were left for character n-gram extractions because some of them could be a result of typos or different word forms. Also, sometimes users do not separate words, e.g "hystericalYou're".

These features were included because they produce good text representation. Particularly, the benefit of character n-grams is that they are robust to inflexion and intentional or accidental typos.

Linguistic: These features are meant to capture stylistic information about tweets and sentiment intensity.

- Number of users mentioning,
- Presence of URL,
- Tweet length,
- Adjectives frequency. As stereotype tweets include more describing words,
- Verbs frequency. As dominance tweets include more verbs
- Punctuation marks frequency,
- Number of repeated punctuation. When the same punctuation mark is repeated more than once, e.g. ??, !!, ?!?!,
- Number of words lengthenings. For instance, "Pleeeeeease",
- Number of capitalized letters,
- Number of emojis,
- Hashtags.

Embedding: Word embeddings capture semantic similarity of words and learn their representations as low-dimensional vectors. Thus, embeddings can give useful information for text categorization. For feasibility, I chose to build an approximated representation of each tweet as an average of word embeddings vectors of all words in a tweet. To avoid adding noise to the representation, stop words were excluded. This approach has been successful in some text mining tasks [9]. Word embeddings were integrated with `gensim` library for python [11] and a public model pre-trained on a Twitter dataset was used [1]. In this model, words are represented by 400-dimensional vectors.

`k-means` algorithm was used to perform clustering. First, I set `k` to 5 for features set selection and to check to what extent categories defined by AMI organizers emerge from the data naturally. Based on the quality of clusters produced by different combinations of features, I chose the features set, which gives the best representation of misogynous tweets. I tried two approaches to evaluate the quality of clusters: internal measures and manual evaluation. External measures, like purity, do not seem to be applicable because the categories in the data set are highly imbalanced and, in most cases, it is hardly possible to determine the correspondence between a cluster and true class. Most cluster members usually were of "discredit" category. The internal measures I tried are Silhouette Coefficient, Calinski-Harabasz Score and Davies-Bouldin Index. These are the common measures for clustering evaluation and they can be applied without additional overhead because the implementation is available in `sklearn` Python library. It turned

[1] https://www.fredericgodin.com/software/

out that these measures do not reflect the good separation of misogyny categories. Changes of measures values when features set was being improved are described in 5 section.

The manual investigation was based on a comparison of a percentage of "true" misogyny categories in clusters with a percentage of these categories in the data set. This approach helped to see if tweets representation with selected features makes it possible to capture "true" categories difference and similarity.

Lastly, I tried to discover new misogyny categories in the data set. A popular **elbow method** for determination of the number of "true" clusters in the data gave controversial results. This is not surprising since the used evaluation measures fail to reflect the good separation of misogyny types. Thus, eventually, I had to choose k based on intuition.

The analysis is implemented in the Python programming language. The source code and results can be found in GitHub repository[2].

5 RESULTS AND ANALYSIS

5.1 Words frequencies in AMI categories

10 most frequent words, excluding stop words, are presented in Table 2. Unsurprisingly, the word "bitch" belongs to top-5 in all clusters. Moreover, for 3 clusters it is the most frequent word. Since it occurs so often, it can be considered a stop word, thus I added it to the list of stop words.

From this statistic, the AMI categories seem sensible. Sexual harassment mostly has words about intercourse or genitals. The most frequent word in the stereotype category is "hysterical", which is often used in discourse about the emotional instability of women. Second and third place belongs to "woman" and "women", which is a sign of generalization in such tweets. Also, "kitchen" occurs often. Six of the top ten words in discredit category are insults. The authors of tweets in the dominance category seem to use "shut the fuck up" phrase and its contraction frequently. "Rape" is the most popular word in derailing class. One more evident characteristic is the frequency of "men" word, which does not occur in the top 10 words in all other categories.

This analysis highlights, that some difference between categories can be captured even with unigrams, thus it should be possible to obtain reasonable clusters for these data.

5.2 Clustering results, k=5

The initial features set I chose for clustering are unigrams and the full list of linguistic features. Although, it produced decent results in terms of internal measures (Silhouette: 0.5, Calinski-harabaz: 4191, Davies-bouldin: 0.59) manual evaluation revealed that the features mainly contributed to the separation of clusters were *tweet length* and *number of capitalized letters*. One cluster contained the tweets written mostly in capital letter and there were 3 clusters, in which tweets were grouped by length: long, average, short. The same clusters were obtained when characters n-grams were used instead of unigrams. This result shows that the semantic content of tweets expressed with tokens or characters n-grams is much less discriminative than these quantitative features. Since tweet length

[2]https://github.com/EvgeniyaMartynova/MisogynyClustering

Table 2: The most frequent words in AMI misogyny categories

Discredit	Sexual harassment	Stereotype	Dominance	Derailing
bitch	bitch	hysterical	bitch	rape
whore	dick	woman	fuck	women
cunt	cunt	women	stfu	men
women	ass	like	shut	woman
hoe	fuck	bitch	women	bitch
fucking	pussy	get	hoe	like
ass	like	hoe	like	want
stupid	whore	know	fucking	whore
skank	rape	#womensuck	get	people
like	fucking	kitchen	pussy	get

or number of capitalized letters is generally unrelated to semantic and could be more informative about the author profile, I removed these features and repeated clustering. After this, internal measures scores significantly worsened (Silhouette: 0.12, Calinski-harabaz: 242, Davies-bouldin: 1.82). One of the clusters contained only one tweet with a large number of punctuation marks. Thus, my next step was to remove *punctuation marks frequency* and *number of repeated punctuation*. The last feature which hampered the discovery of misogyny categories was *number of users mentioning*. When the linguistic features mentioned before were removed, one of the clusters mostly contained tweets with two or more users names. This feature was removed as well.

Bigrams and trigrams gave very similar results to unigrams, therefore I did not include them into further analysis. Overall, unigrams alone or with the set of selected linguistic features did not produce meaningful clusters.

Interesting results were obtained with character n-grams. Trigrams solely produced a cluster with 169 tweets 144 of which had stereotype category. Another cluster had 370 tweets and 54 of them were of derailing category, which is more than half of all tweets in this category and its share in this cluster is 0.146, which is three times higher than in the whole data set. The largest cluster (1252 tweets) had the highest share of discredit and sexual harassment tweets (0.59 and 0.21). Character 4-grams gave similar results, although the separation of categories was less clear, and 5-grams performed poorly producing 3 clusters with one message only. Finally, character 3- and 4-grams with linguistic features selected on the previous step produced one more interpretable cluster with high shares of dominance and stereotype tweets (0.22 and 0.23).

Despite the loss of information due to averaging of vectors, word embeddings features had some discriminative power as well. When used solely, they generated a cluster similar to the one with stereotype majority (431 tweets overall and 160 stereotype tweets) and also there were two clusters with clear dominance (35 of 52) and discredit (187 of 250) majority. Other two clusters were similar to the produced with character n-grams.

Based on these results, I decided to include unigrams, character 3- and 4-grams and word embeddings into the final test. Also, I experimented with different sets of linguistic features apart from already removed. After looking into the results I found that unigrams,

character 3- and 4-grams, word embeddings, adjectives and verbs frequencies set of features produces the most meaningful set of clusters. The structure of the clusters is given in Table 3 in the Appendix. According to evaluation measures, these clusters are overlapping and not separated well. Silhouette score: 0.05, Calinski-harabaz: 43.5 and Davies-bouldin: 5.13.

It is challenging to derive a reliable conclusion from these results. Apparently, stereotype and derailing categories can be most easily discriminated. Cluster 3 allows suggesting that there is a similarity between some sexual harassment and dominance tweets. However, since this cluster is small, it could be explained differently. Lastly, sexual harassment category looks the most similar to discredit.

5.3 Mining new categories

If we want to mine more misogyny categories from the given data set, two option are possible. First, we can assume that tweets from unobserved categories are distributed across all annotated categories. In this case, it is reasonable to cluster the whole data set. Alternatively, the discredit category can be clustered. It is the largest and the most vaguely defined category, so it might contain subcategories. I decided to try both approaches.

Since `k-means` algorithm is used for clustering, a reasonable number of clusters should be determined. It is common to select it with the elbow method by checking how the value of a function that evaluates clustering changes for different k. The results of this analysis are depicted in Figure 3 in the Appendix. The measures discussed at the end of the section 4 and SSE (inertia) calculated by `sklearn k-means` implementation were used. From the graphs, it is clear that with the extracted features neither k produces good clusters in terms of these metrics.

Hence, I decided to select k based on the intuition. As for the whole data set, [1] describes two more categories that were merged into others in the final set: *objectification* and *threats of violence*. Also, with $k = 5$, clusters with one tweet only were produced quite often, therefore I do not expect the data set to contain much more new categories and assume that it could be possible to identify three more categories. Thus, I set k to 10.

For clustering of discredit category, I set k to 6. Considering the number of tweets in this category, it is unlikely that it contains more subcategories. And from my previous experience with clustering, the most robust clusters are kept when k is increased, so if there are clear subcategories they will not be missed this way.

To evaluate the clusters and understand which known or unobserved category they may be related to, I sampled and read 50 tweets from each cluster. Annotation was not included in samples to reduce bias.

5.3.1 All misogynous tweets results. Three clusters contained only a few tweets (1, 2 and 6), which look like outliers. A single tweet consists of the repetition of "You're hysterical". Six were all rather long and contained almost only swear words. For the two tweets, the similarity is less clear, but they were repetitive as well.

Three more clusters could be roughly related to separate categories and it was straightforward to spot tweets similarity:

- **Stereotype** cluster (159 tweets). All tweets in this cluster contained "hysterical" word at least once. The goal of these tweets was to insult or discredit women by referring to stereotypes. This cluster is almost equal to stereotype cluster obtained in section 5.2.
- **Slut-shaming** cluster (214 tweets). Quite a lot of tweets insulted women because of (imaginary) choices in personal life, however not all tweets in the cluster are slut-shaming. All tweets in this cluster contained "whore" word at least once.
- **Rape justification** cluster (505 tweets). This is the least pure cluster, however, a sample contained quite a few tweets with rape or violence justification. The word "rape" occurs more frequently in this cluster than in others. Also, it has 51 of all 103 derailing tweets, which is likely to have this subcategory.

For the other four clusters, I did not manage to spot clear categories. From the analysis above it is clear that the model tends to group tweets mostly based on words occurrences and does not reflect semantic well. Only such quantitative similarities were noticeable: e.g. one cluster had mainly long tweets and many "fuck" occurrences, another had only short tweets and abundance of sex-related words. However, the semantic content of tweets in the sample was quite different.

Eventually, two new misogyny categories were discovered in the data set: **slut-shaming** and **rape justification**. Rape justification has been briefly discussed in derailing category definition in [1], whereas slut-shaming was not mentioned by the creators of the data set at all. Table 4 in Appendix contains examples of tweets from these categories.

5.3.2 Discredit category results. With manual evaluation, I could not determine any meaningful subcategories in the produced clusters apart from the slut-shaming cluster. In the slut-shaming cluster, generated with all tweets, 153 of 214 samples had discredit category and the corresponding discredit cluster had 150 tweets. This shows the robustness of this cluster.

6 DISCUSSION AND OUTLOOK

In this paper, I examined misogyny categories present in AMI task data set. The performed analysis demonstrates that studying of hate speech in social media with NLP methods is a challenging and highly non-trivial task. In particular, we saw how difficult the choice of the representative feature set is. And even with the best features set semantic content is not captured.

Nevertheless, analysis of misogynous tweets clustering with $k = 5$ revealed that even with imperfect features set some AMI categories and similarities between categories could be determined in an unsupervised manner.

Although most of the clusters generated for mining of the new categories were senseless, two new categories were discovered: **slut-shaming** (which is also a subcategory of discredit class) and **rape justification**.

The main future work direction I see is devising a more complex framework for features extraction. While features set fails to extract semantic content of a tweet, the task will remain mostly manual and time-consuming. Also, capturing of tweets semantic content might not be possible without information about the context. Thus, in future, we should consider building data sets which contain not single tweets but threads and adjacent information.

REFERENCES

[1] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*. Springer, 57–64.

[2] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 45–54.

[3] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 54–63. https://doi.org/10.18653/v1/S19-2007

[4] James V Check and Neil M Malamuth. 1983. Sex role stereotyping and reactions to depictions of stranger versus acquaintance rape. *Journal of Personality and Social psychology* 45, 2 (1983), 344.

[5] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

[6] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian* 12 (2018), 59.

[7] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018.. In *IberEval@ SEPLN*. 214–228.

[8] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–30.

[9] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.

[10] Bailey Poland. 2016. *Haters: Harassment, abuse, and violence online*. U of Nebraska Press.

[11] Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

[12] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. http://www.aclweb.org/anthology/N16-2013

[13] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666* (2019).

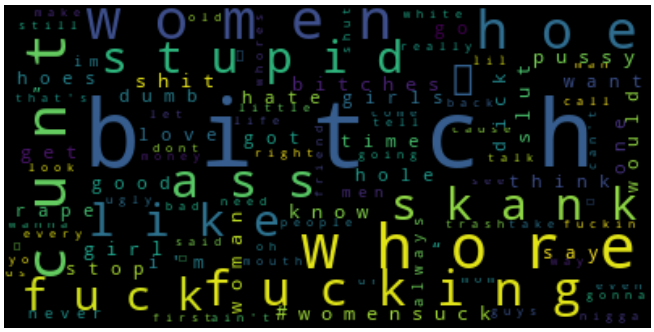
[14] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983* (2019).

APPENDIX

Table 3: The structure of clusters generated with the final set of features. The percentage is given in the brackets.

	Overall	Discredit	Sexual harassment	Stereotype	Dominance	Derailing
1	165	12 (0.07)	3 (0.02)	140 (0.85)	7 (0.04)	3 (0.02)
2	1098	645 (0.59)	204 (0.18)	97 (0.09)	126 (0.11)	26 (0.02)
3	107	56 (0.52)	25 (0.23)	4 (0.04)	22 (0.21)	0 (0)
4	271	120 (0.44)	31 (0.11)	34 (0.13)	41 (0.15)	45 (0.17)
5	604	322 (0.53)	133 (0.22)	44 (0.07)	76 (0.13)	29 (0.05)

Figure 1: Word clouds for AMI categories



Discredit

Sexual Harassment

Stereotype

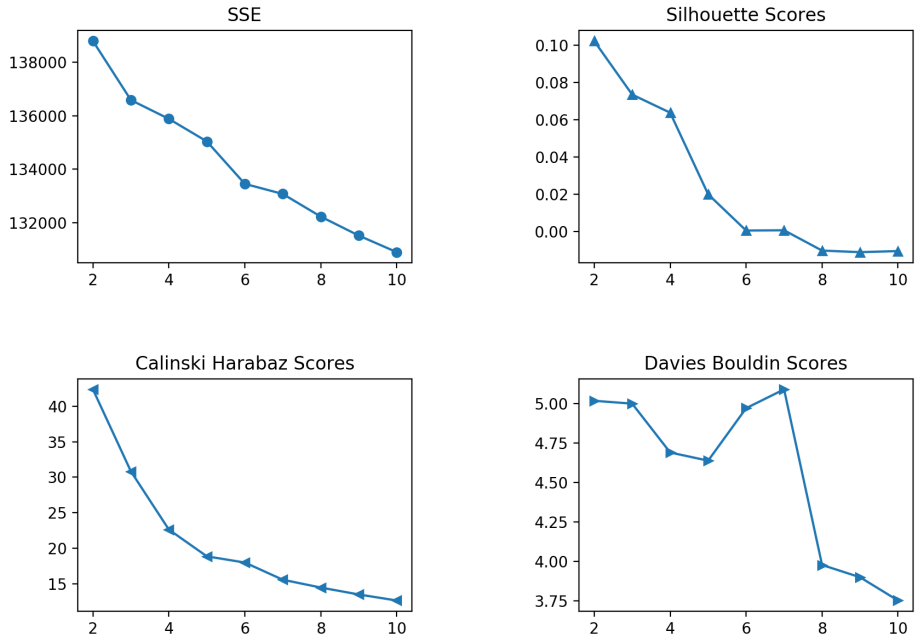
Dominance

Derailing

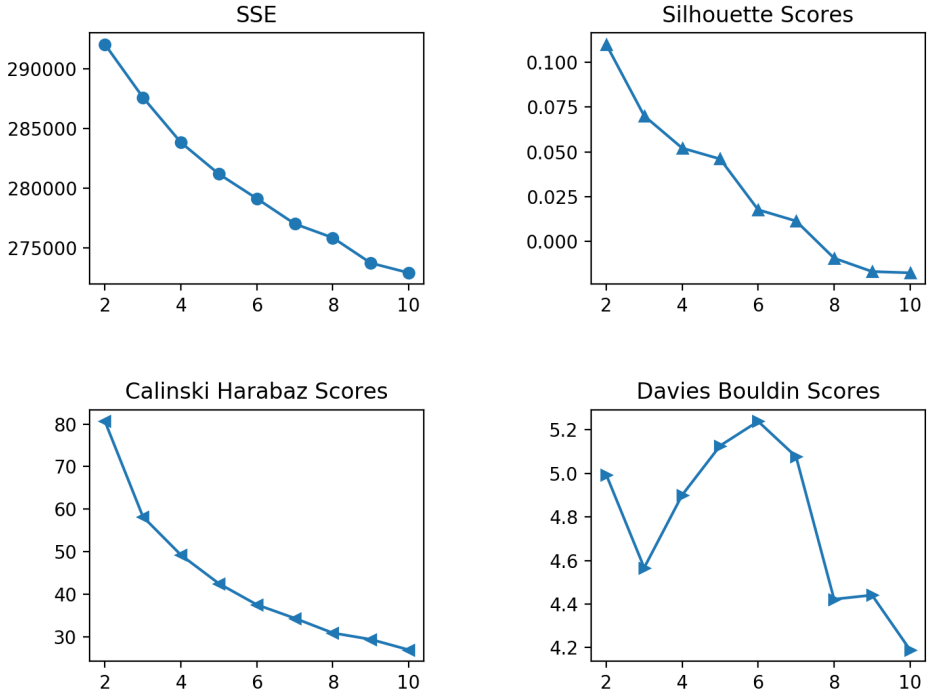
Table 4: Examples of tweets from the new categories

Slut-shaming
But remember, you cheated on me first, you're a whore, you're a whore, this is whore.
#MeToo women always blame men for everything . How about not dressing like a whore?
There is a big reason you are lonely and it's because you are such a fricken whore that no man can trust or respect you, so stop bitchin'.
Rape justification
This isn't rape, or anything like it. This a crude pass. This is not 'rape culture.' IT'S WHORE CULTURE.
Remember, this the same bitch that was on FOX news saying rape culture was an attack on young men. Its her pussy be…
Powerful men in Hollywood have been preying on women because that's how lots of men are built ... so says Dick Wolf.

Figure 3: Clustering analysis with different k



Discredit tweets



All misogynous tweets