Predicting mental disorders from transcripts

By Ludvig Renbo Olsen, 201507847
Natural Language Processing
Cognitive Science, Aarhus University

Abstract

This study extends (Olsen, 2018), where we diagnosed mental disorders from speech recordings, by attempting to diagnose schizophrenia and depression from the transcripts of those same recordings. Using multiclass classification, we achieve a higher than chance overall accuracy (0.547, downsampled dataset). We further present a set of metrics for finding the words that are used differently between the diagnoses. Code for calculating these is shared as an R package.

The project code can be found at https://github.com/LudvigOlsen/nlp_exam (project and modelling) and https://github.com/LudvigOlsen/vocabular2 (word usage metrics, R package).

Ludvig Renbo Olsen
201507847

Introduction

Mental disorders like depression and schizophrenia are difficult to diagnose due to their symptom heterogeneity (Cuthbert & Insel, 2013; Fried & Nesse, 2015). By finding biomarkers of the disorders, we can better detect, understand and support patients (Singh & Rose, 2009). In this exploratory study, we tackle the problem in two ways. First, we develop metrics for finding words that are uniquely applied by one of the patient groups. Then, we perform multiclass classification to discriminate between the three classes: depressed, schizophrenic and control.

Dataset

The dataset contains transcriptions from 402 Danish speaking subjects (224 controls, 67 depressed (first episode or chronic), 111 schizophrenics (first episode)) from 9 datasets (Bang, 2009; Bliksted et al., 2014, 2018, 2017; Ladegaard et al., 2014; unpublished studies by Line Gebauer, Emma Fowler, Heine Lund Pedersen, and Vibeke Bliksted). The subjects, ranging between 18 - 67 years of age, performed up to 10 trials of the Frith-Happé animations task (Abell et al., 2000), which was developed to assess the ability to ascribe mental states (theory of mind). Table 1 contains diagnosis-wise statistics.

Diagnosis	**# Subjects**	**Avg. # trials**	**Avg. # chars per trial**
Control	224	8.31	242
Depression	67	8	152
Schizophrenia	111	7.77	229

Table 1: Diagnosis-wise statistics on the number of subjects, average number of trials, and average number of characters per trial.

Methods

We perform two types of analyses: First, we iteratively develop a set of metrics for finding words that are uniquely used by a diagnosis[1]. Secondly, we perform multiclass classification in order to assess the discriminability of the diagnoses based on their word usage.

Finding differences in vocabulary

To assess the difference in word usage between the diagnoses, we calculated a set of possibly novel term-frequency metrics, similar to the well-known TF-IDF metric. Where the TF-IDF compare the current document (i.e. diagnosis) to the entire corpora (including the current document), we instead create one-vs-rest comparisons, such that the current document is compared to all documents but itself. This leads us to the TF-IRF (Term

[1] The results of this analysis are in the appendix.

Frequency - Inverse Rest[2] Frequency). We find that TF-IRF and TF-IDF are strongly correlated (cor = 0.999989), why we interpret them similarly. The following equations show the differences between TF-IDF and TF-IRF. As we see, they are so similar, that it would usually make most sense to calculate the TF-IDF, as we only need to calculate the IDF once, whereas IRF needs to be calculated for each document. To be consistent with our other metrics though, we stick with TF-IRF for this paper.

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t'}^{d} f_{t',d}}$$

$$idf(t,D) = log\frac{|D|}{1+|\{d \in D : t \in d\}|}$$

$$irf(t,d,D) = log\frac{|D|-1}{1+|\{d' \in D : t \in d' \wedge d' \neq d\}|}$$

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$$

$$tfirf(t,d,D) = tf(t,d) \cdot irf(t,d,D)$$

Equations 1-5: TF (term frequency), IDF (inverse document frequency), IRF (inverse rest frequency), TF-IDF and TF-IRF. $\{d' \in D : t \in d' \wedge d' \neq d\}$ denotes the documents (excluding *d*) that contain *t*. $f_{t,d}$ denotes the count of *t* in *d*.

In our experiments, both TF-IDF and TF-IRF tend to simply select the terms with the highest term frequency that are only in the current document. Both are highly negatively correlated with the term frequency (cor < -0.99). This may be an effect of having only 3 “documents”. To better find the words that are used differently within a diagnosis, we thus present the three metrics TF-RTF, TF-NRTF and Relative TF-NRTF.

TF-RTF (Term Frequency - Rest Term Frequency)

The RTF (Rest Term Frequency) is the sum of the term frequencies in the other documents. This is subtracted from the term frequency in the current document. TF-RTF is thus only positive when the term frequency is higher in the current document than the term frequencies in the rest of the corpus combined. We further chose to set negative values to 0 to fully focus on those words, but this step is not required and not described in the formulas.

$$rtf(t,d,D) = \sum_{d' \neq d}^{D} tf(t,d')$$

$$tfrtf(t,d,D) = tf(t,d) - rtf(t,d,D)$$

Equations 6-7: RTF (Rest Term Frequency) and TF-RTF.

[2] We did not spend our time finding the optimal names for the metrics.

Given a large corpora, there might not be any words that get a positive TF-RTF score. This leads us to the TF-NRTF metric.

TF-NRTF (Term Frequency - Normalized Rest Term Frequency)

As our selected TF function ensures that the frequencies add up to 1 document-wise, the NRTF (Normalized Rest Term Frequency) is simply the average TF in the other documents, instead of the sum as in RTF. This means, that both the TF and NRTF terms will have a value between 0 and 1. By averaging the RTFs, every document contributes equally to the rest distribution. This can be useful when working with imbalanced datasets. If we get a positive value, the term is used more frequently in this document than on average in the other documents. For some use cases, the Maximum Rest Term Frequency (MRTF) could also be useful as it would tell us whether the term is used the most in this document. We won't go further into that here though.

$$nrtf(t, d, D) = \frac{rtf(t, d, D)}{|D| - 1}$$

$$tfnrtf(t, d, D) = tf(t, d) - nrtf(t, d, D)$$

Equations 8-9: NRTF (Normalized Rest Term Frequency) and TF-NRTF.

TF-RTF and TF-NRTF had a correlation coefficient of 0.427. TF-RTF was not correlated with TF (cor = 0.07), while TF-NRTF had a fairly high correlation with TF (cor = 0.61).

Rel TF-NRTF (Relative Term Frequency - Normalized Rest Term Frequency)

In TF-NRTF, the frequently used words are almost bound to end up in the top TF-NRTF scores for one of the documents. What we might actually be interested in is the relative distance between the current document and the average rest population. This, on the other hand, would likely be dominated by the very infrequent words, as the difference between 0.0003 and 0.001 is the same as 0.03 and 0.1 and likely happens more often. As the latter seems like a more interesting difference, we multiply the relative difference with the TF:

$$\varepsilon(t, d, D) = 1 \,/ \sum_{d' \neq d}^{D} f_{t,d'}$$

$$rel\ tfnrtf(t, d, D) = tf(t, d)^{\beta} \frac{tfnrtf(t, d, D)}{log(1 + nrtf(t, d, D) + \varepsilon(t, d, D))}$$

Equations 10-11: Rest Epsilon and Relative TF-NRTF.

Epsilon (ε) is added to avoid zero-division. It is calculated to resemble +1 smoothing in the rest population. The beta (β) exponentiator allows us to control the influence of the term frequency. By setting it to 0, we get the relative difference (log scaled) (TF-NRTF cor = 0.07, rank cor = 0.81). Figure 1 shows the correlation at different beta values. We use β=1 (TF-NRTF cor = 0.75, rank cor = 0.98).

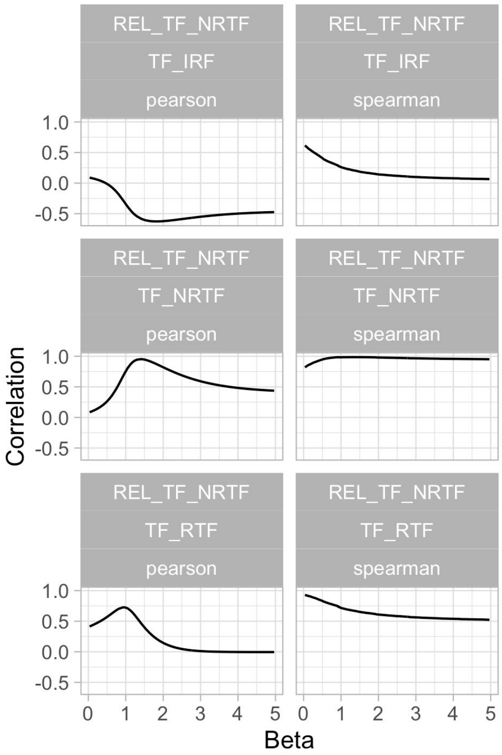


Figure 1: Correlation coefficients for different beta values in Relative TF-NRTF with the three other metrics.

Appendix B contains word clouds for each of the 4 metrics, along with short discussions. Appendix C contains 4 tables with the 10 highest scores for each diagnosis, for each metric. The metrics were calculated on the preprocessed (see *Preprocessing*) dataset.

Discriminating between the diagnoses from word usage

We used a set of multiclass classifiers to discriminate between the diagnoses from the word usage in a trial (bag of words). This section describes the dataset preprocessing, choice of model architectures and the model selection and evaluation architecture.
Note, that while we did fine-tune a BERT transformer model[3] on the task, it did not perform better than the simpler model architectures and was left out due to time and character constraints.

Preprocessing

To reduce the influence of the individual transcribers' habits, we sought to bring the transcripts to a common format. The following steps were performed in python 3.7.4: 1) Some transcribers used a forward slash (/) to symbol a break in the speech. These were replaced with dots, which were also commonly used. 2) Hyphens that were not in-between two words were removed. 3) Transcribers sometimes added extra info in parentheses, why we removed the parentheses and their content. 4) For some of the transcriptions, the special Danish characters æ,ø,å had been replaced by ae,oe,aa respectively. We converted them back to æ,ø,å. 5) Any character not matching the following regular expression was removed:

[3] We fine-tuned the pretrained BERT base-multilingual-cased from Hugging Face from https://github.com/huggingface/transformers
We added two dense layers with 512 and 3 units based on multiple experiments.

[^A-Za-zæøåÆØÅ.,\-!?]. 6) Multiple consecutive whitespace characters were replaced with a single whitespace. 7) Stopwords and punctuation marks were removed. 8) Words were lower cased.

We further added subwords to the classification dataset to see if they would increase the performance. These were every 2 and 3 letter windows of the words and were prefixed with "##" (e.g. "røde" would become "##rø", "##ød", "##de", "##rød", and "##øde", along with the actual word).

We did not use lemmatization or stemming for the current results. The subwords will likely have brought some of the same information to the classifiers.

The classifiers were trained on bag of words with token counts by default. We trained on three different n-gram upper limits (1, 2, and 3) to see if adding larger n-grams would increase performance. Subwords were considered single grams like the words.

Balancing the dataset

As the dataset is highly imbalanced, we created up and downsampled versions and trained the classifiers on all three versions for comparison.

Model architectures

Using scikit-learn (Pedregosa et al., 2011), we fitted a Support Vector Machine (SVM; LinearSVC) and three types of Naïve Bayes (NB) classifiers (ComplementNB, BernoulliNB, and MultinomialNB), all with default settings. Later work should tune the various hyperparameters. As mentioned previously, we also fine-tuned a pre-trained BERT transformers model, but excluded it due to a lack of performance improvements.

Cross-validation

In order to run 10-fold subject-wise, stratified cross-validation, we split the dataset in 10 folds using the R package groupdata2 (Olsen, 2019, p. 2). As running repeated cross-validation would be too time-consuming, we optimized for similarity between the folds, in the hope of more stable fold results. We created 20 unique fold columns (each being a 10-fold split) stratified by diagnosis and numerically balanced by the length of sentences (number of characters), while ensuring that each subject was placed in only one of the folds. From these 20 splits, we selected one with relatively low variation between the folds in the number of transcripts, the number of transcripts per diagnosis, and the number of characters. This was done for each version of the dataset (upsampled, downsampled, imbalanced).

Baseline evaluations

For our results to be useful, they must at least be better than random guessing. Usually, the chance-level is found analytically. F.i., as we have 3 classes, there's a ⅓ chance of predicting each of the classes, and so we should expect an overall accuracy of 0.33. In practice though, we can get a higher accuracy than that, especially on smaller datasets. Hence, we take a different approach and evaluate 100 sets of random probabilities with the *baseline* function in the R package cvms (Olsen & Zachariae, 2019).This tells us the expected value for each metric (which should be very close to the analytical chance level) along with a standard deviation and the minimum and maximum values obtained.

For imbalanced datasets, there's a tendency for some classifiers to always predict the majority class, in order to quickly minimize the loss function. Hence, we also evaluate a set of "all controls" predictions.

Results

In order to discuss the three balancing approaches (Upsampling, Downsampling, Imbalanced), we report the results for each[4], along with their baselines. We use the macro balanced accuracy metric as our selection criterion, due to the inclusion of the imbalanced dataset, but also report overall accuracy, macro F1 score, macro precision, and macro recall. For the imbalanced dataset, we further provide the baseline evaluation when always predicting "Control".

Imbalanced dataset

On the imbalanced dataset, the Bernoulli Naïve Bayes model with subwords achieved the highest balanced accuracy of 0.592 (baseline = 0.5). With an overall accuracy of 0.534 (baseline = 0.333) and an F1 score of 0.451 (baseline = 0.298), it performs well above chance. The model could have achieved a higher overall accuracy by always predicting "Control", although that would have decreased the balanced accuracy to 0.5. Table 2 shows the best results for each of the model architectures.

Model	**N Grams**	**Overall Accuracy**	**Balanced Accuracy**	**F1**	**Precision**	**Recall**
BSL Mean (SD)		0.333 (0.00532)	0.500 (0.00504)	0.298 (0.00519)	0.333 (0.00493)	0.334 (0.00740)
BSL Max.		0.346	0.515	0.312	0.346	0.357
BSL All Control		0.612	0.5	NaN	NaN	0.333
SVM	1	0.552	0.570	0.436	0.442	0.432
SVM SW	1	0.540	0.563	0.425	0.429	0.423
NB Multinomial	1	0.587	0.565	0.425	0.463	0.418
NB Multinomial SW	1	0.521	0.588	0.445	0.437	0.460
NB Bernoulli	1	0.588	0.545	0.388	0.442	0.390
NB Bernoulli SW	1	0.534	**0.592**	0.451	0.444	0.464
NB Compl.	1	0.538	0.575	0.432	0.434	0.432
NB Compl. SW	1	0.518	0.570	0.423	0.422	0.429

Table 2: Best results for each of the model architectures on the imbalanced dataset, with and without subwords (SW). The *BSL* rows are the baseline evaluations.

[4] Due to the character constraint, the results for the upsampled dataset can be found in Appendix D.

Figure 2 shows the confusion matrix. Due to the class imbalance, we focus on the column and row percentages. For the first tile (schizophrenia / schizophrenia), the column percentage tells us how often the model predicts schizophrenia when the true class is schizophrenia (42.3% of the time). The row percentage tells us how often the model is correct when predicting schizophrenia (39.9% of the time). In the “correct” diagonal, the column percentages are the class-level recall scores, while the row percentages are the class-level precision scores. Importantly, the model doesn’t always predict the majority class.

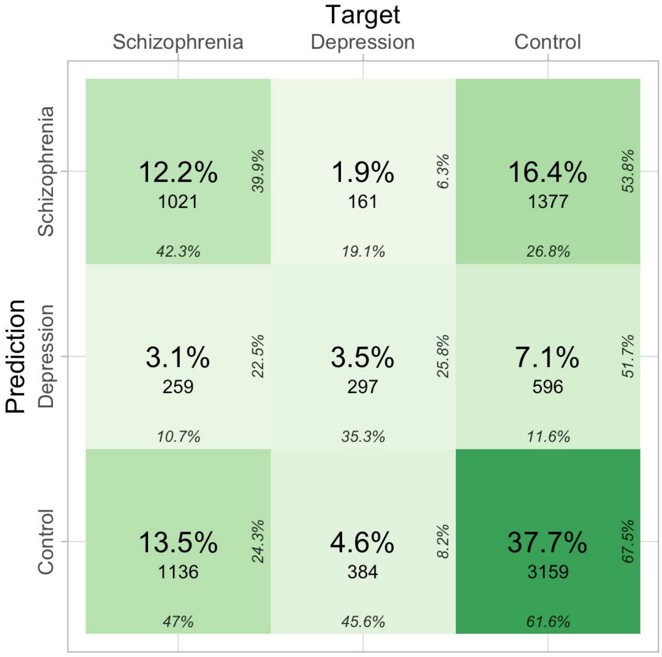


Figure 2: Confusion matrix for the best model on the imbalanced dataset. Each tile contains the count, the normalized count (overall percentage), the column percentage (bottom) and the row percentage (right). The color intensity is based on the count.

61.6% of the controls are correctly identified, while 26.8% are predicted to be schizophrenic and 11.6% to be depressed. 35.3% of the depressed patients are predicted to be depressed, with 45.6% to be controls and 19.1% to be schizophrenic. 42.3% of the schizophrenics are correctly identified, with 47% being predicted to be controls and 10.7% depressed.

Downsampled dataset

On the downsampled dataset, the support vector machine achieved the highest balanced accuracy of 0.660 (baseline = 0.5). With an overall accuracy of 0.547 (baseline = 0.334) and an F1 score of 0.539 (baseline = 0.334), it performs well above chance. Table 3 shows the best results for each of the model architectures.

Model	**N Grams**	**Overall Accuracy**	**Balanced Accuracy**	**F1**	**Precision**	**Recall**
BSL Mean (SD)		0.334 (0.00961)	0.500 (0.00720)	0.334 (0.00960)	0.334 (0.00959)	0.334 (0.00961)
BSL Max.		0.358	0.518	0.358	0.358	0.358
SVM	1	0.547	**0.660**	0.539	0.543	0.547
SVM SW	1	0.503	0.627	0.501	0.500	0.503
NB Multinomial	1	0.527	0.645	0.526	0.525	0.527
NB Multinomial SW	1	0.485	0.614	0.485	0.485	0.485
NB Bernoulli	1	0.538	0.653	0.528	0.529	0.538
NB Bernoulli SW	1	0.515	0.636	0.508	0.507	0.515
NB Compl.	1	0.511	0.634	0.508	0.507	0.511
NB Compl. SW	1	0.478	0.609	0.475	0.474	0.478

Table 3: Best results for each of the model architectures on the downsampled dataset, with and without subwords (SW). The *BSL* rows are the baseline evaluations.

Figure 3 shows the confusion matrix. 45.7% of the controls are correctly identified, with 26.1% predicted to be schizophrenic and 28.1% to be depressed. This time, 74.9% of the depressed patients are predicted to be depressed, with 16.5% of them predicted to be controls and 8.6% of them schizophrenic. 43.5% of the schizophrenics are correctly identified, with 32.9 being predicted to be controls and 23.6% depressed. Compared to the imbalanced dataset, only the minority class (Depression) seems to gain from the downsampling.

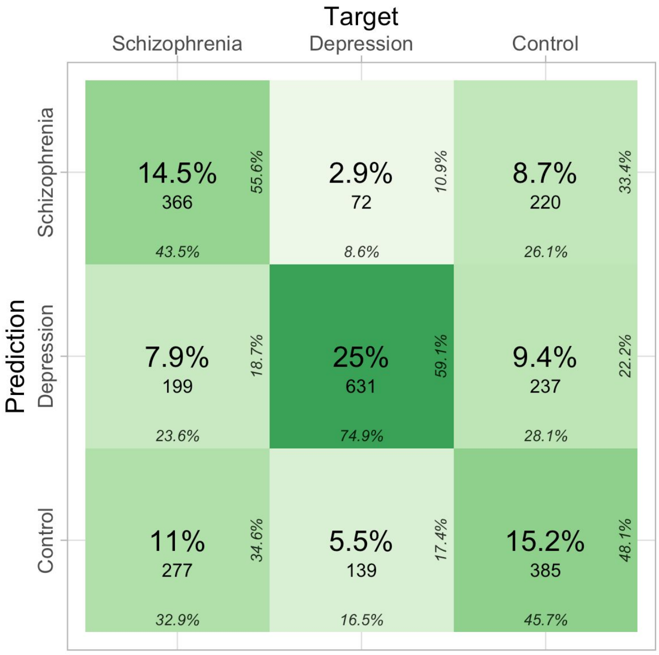


Figure 3: Confusion matrix for the best model on the downsampled dataset.

Discussion

We have discriminated between the three diagnoses with higher than chance performance on each dataset (imbalanced, upsampled, downsampled; with and without subwords). This indicates that the participants used different words to describe the triangle animations. Whether or not these differences were caused by the experimental setup or are inherent to the diagnoses, our results call for further exploration. Extending the analyses from differences in word usage (bag of words) to sentence structures and emotional content might tell us more about how the patients process the stimuli. We would also like to see larger and more diverse datasets being collected. Currently, the subjects are not diverse enough for the algorithms to be useful in practice.

While we applied a set of model architectures to the task, there are plenty of classifiers to try and hyperparameters to tune. The current results should thus be regarded as simple baselines, not state of the art[5] classifiers.

We developed a set of metrics (TF-IRT, TF-RFT, TF-NRFT, and Relative TF-NRFT) for finding the words that are used differently by one of the patient groups. While we haven't quantified their meaningfulness, we so far find them theoretically meaningful. We consider them rough ideas of possibly useful quantifications for comparing vocabularies. Going forward, they require more thinking, better naming, and comparison to other available metrics.

Ethics statement: While having access to quick and cheap diagnostic tools can help doctors and patients tremendously, it also increases the risk of currently private information becoming public. Most people have supplied substantial amounts of writing to various internet services, which could be used by companies and governments to create mental profiles of users/citizens for financial and political gains. While the current study doesn't appear to increase that risk significantly, it is worth considering going forward.

[5] We realize it's incredibly rare for a paper not to claim SOTA results in machine learning based NLP.

References

Bang, D. (2009). *Pronomial use in ASD. [Title not confirmed]* [Bachelor]. Aarhus University.

Bliksted, V., Fagerlund, B., Weed, E., Frith, C., & Videbech, P. (2014). Social cognition and neurocognitive deficits in first-episode schizophrenia. *Schizophrenia Research*, *153*(1), 9–17.

Bliksted, V., Frith, C., Videbech, P., Fagerlund, B., Emborg, C., Simonsen, A., Roepstorff, A., & Campbell-Meiklejohn, D. (2018). Hyper-and Hypomentalizing in Patients with First-Episode Schizophrenia: FMRI and Behavioral Studies. *Schizophrenia Bulletin*.

Bliksted, V., Videbech, P., Fagerlund, B., & Frith, C. (2017). The effect of positive symptoms on social cognition in first-episode schizophrenia is modified by the presence of negative symptoms. *Neuropsychology*, *31*(2), 209.

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, *11*(1), 126.

Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR* D study. *Journal of Affective Disorders*, *172*, 96–102.

Ladegaard, N., Lysaker, P. H., Larsen, E. R., & Videbech, P. (2014). A comparison of capacities for social cognition and metacognition in first episode and prolonged depression. *Psychiatry Research*, *220*(3), 883–889.

Olsen, L. R. (2018). *Automatically diagnosing mental disorders from voice* [Bachelor]. Aarhus University.

Olsen, L. R. (2019). *groupdata2: Creating Groups from Data*. https://CRAN.R-project.org/package=groupdata2

Olsen, L. R., & Zachariae, B. H. (2019). *cvms: Cross-Validation for Model Selection*. https://github.com/ludvigolsen/cvms

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Singh, I., & Rose, N. (2009). Biomarkers in psychiatry. *Nature*, *460*(7252), 202.

Appendix

Appendix A: Metrics by rank

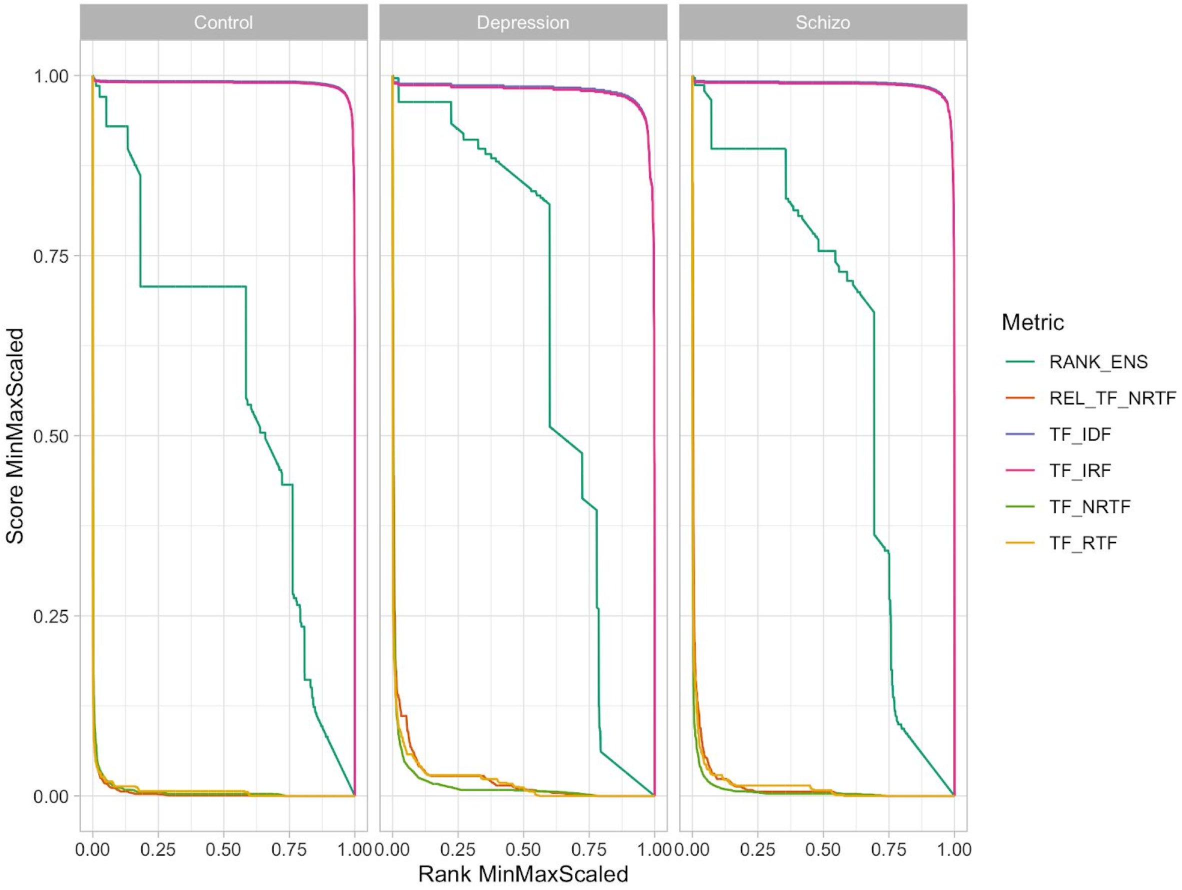
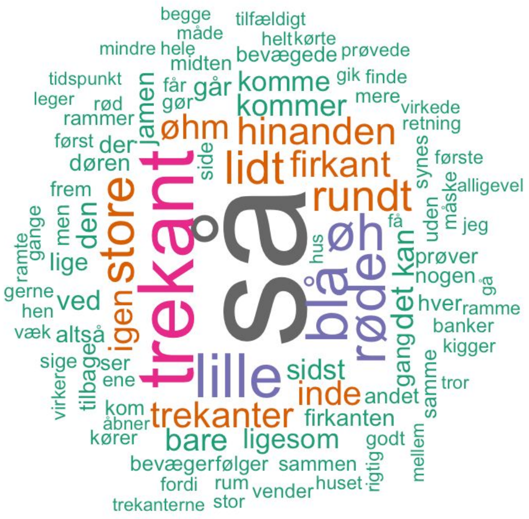


Figure A1: MinMax scaled score by MinMax scaled rank for six different metrics.
The rank ensemble is the ranked sum of ranks from the other metrics (excl. TF-IDF). I.e. RANK_ENS = rank(rank(TF-IRF) + rank(TF-RTF) + rank(TF-NRTF) + rank(Rel TF-NRTF)).
Whether these are meaningful is not for an appendix to address!

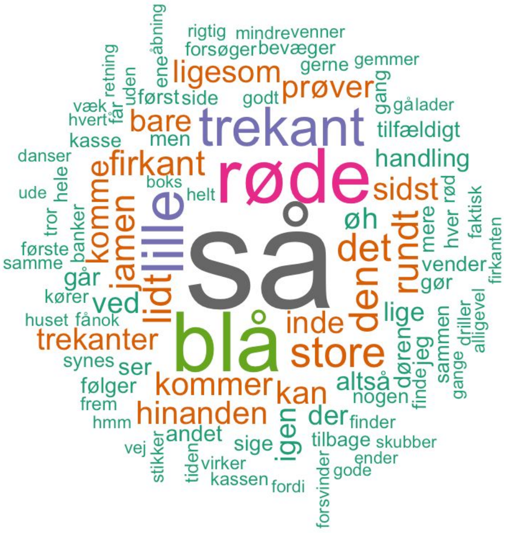
Ludvig Renbo Olsen
201507847

Appendix B: Word clouds by metrics

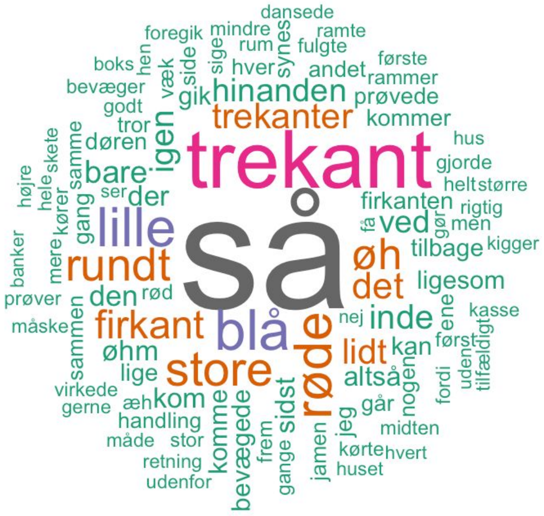
TF (Term Frequency)



Control



Depression

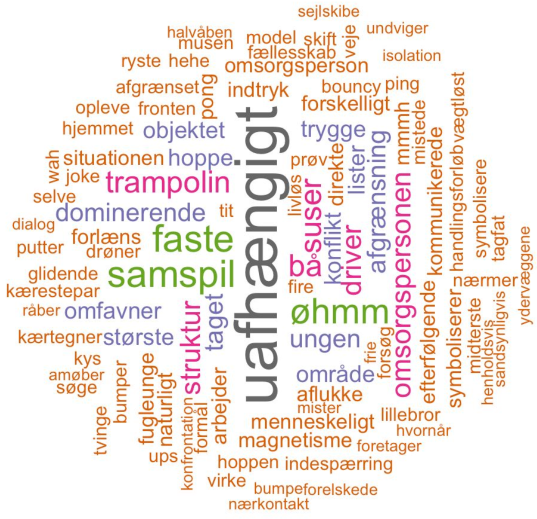


Schizophrenic

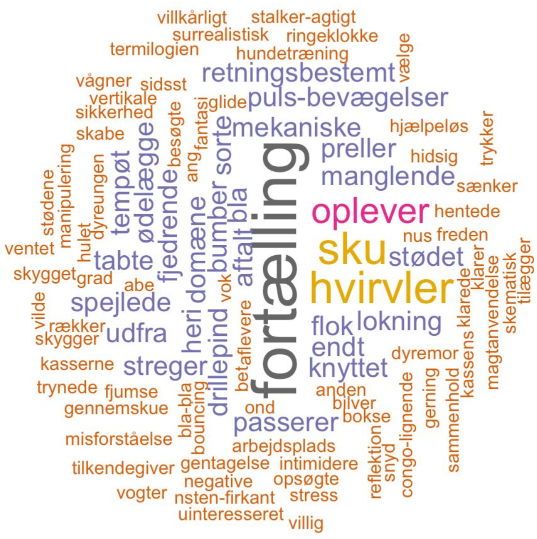
Figure A2: Term Frequency word clouds.

The term frequency word clouds leaves us with the impression that the three participant groups mostly use the same words. The highest scoring word “så” means both “so” and “saw”. “Trekant” means “triangle”, while “røde” and “blå” loosely means “the red one” and “the blue one” respectively.

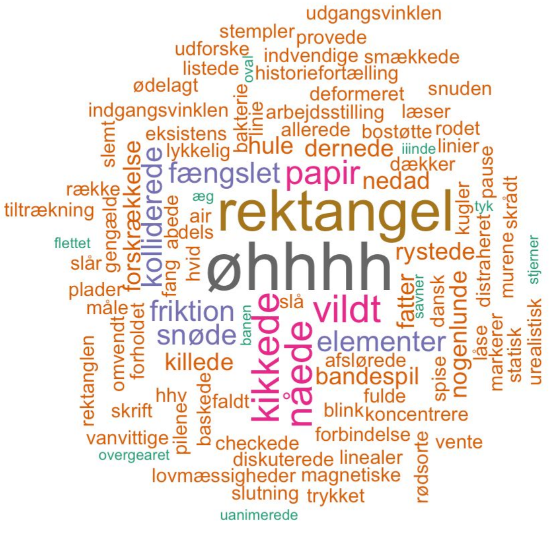
TF-IRF (Term Frequency - Inverse Rest Frequency)



Control



Depression

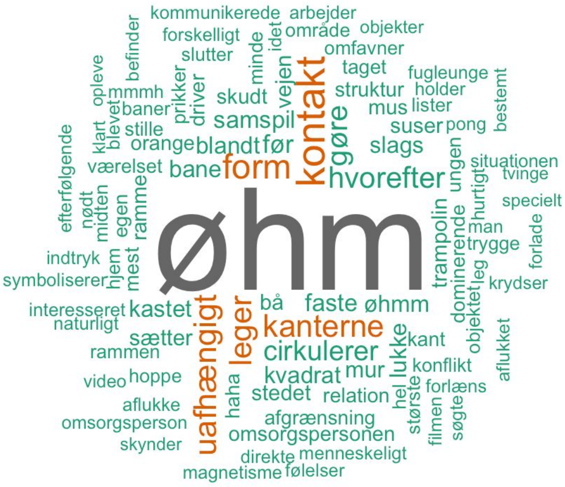


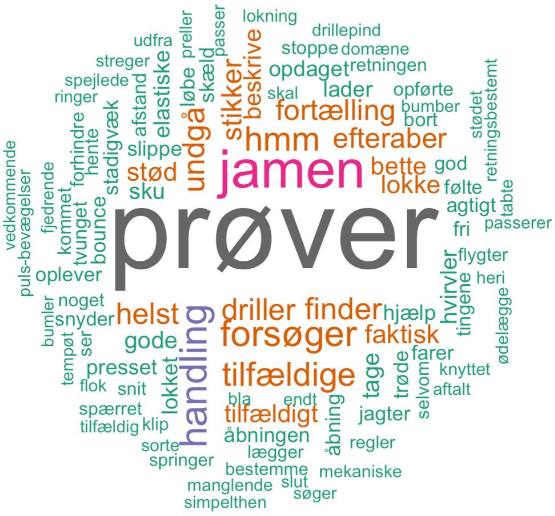
Schizophrenic

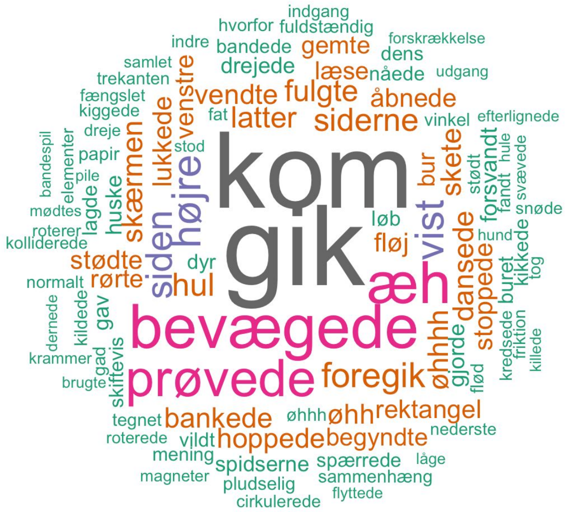
Figure A3: Term Frequency - Inverse Rest Frequency word clouds.

Each of the highly scoring words are only used by the subjects in that class. They are not necessarily very frequent. The highest scoring and most frequent word in the Control cloud “uafhængigt” means “independent” and was used 21 times in total. For Depression, the word “fortælling” means “narration” and is used 6 times in total. For schizophrenic, the word “øhhhh” is similar to “ehm” and was used 12 times in total. It is only the number of h’s that makes it special in this case though, as “øh” and “øhm” were both commonly used. Here, a collapsing of similar words, f.i. with lemmatization, might increase the relevance of words found.

TF-RTF (Term Frequency - Rest Term Frequency)







Control

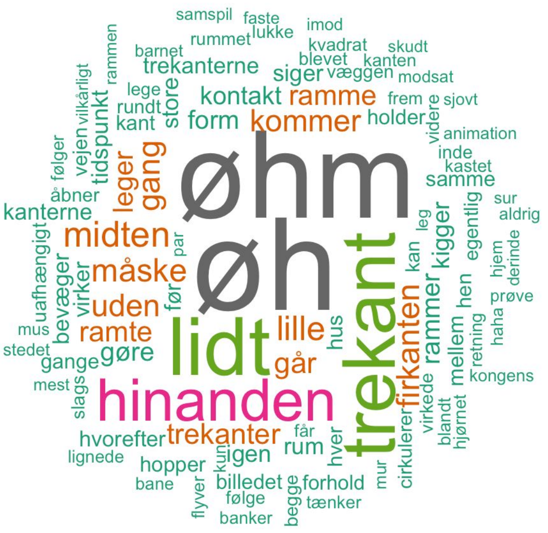
Depression

Schizophrenic

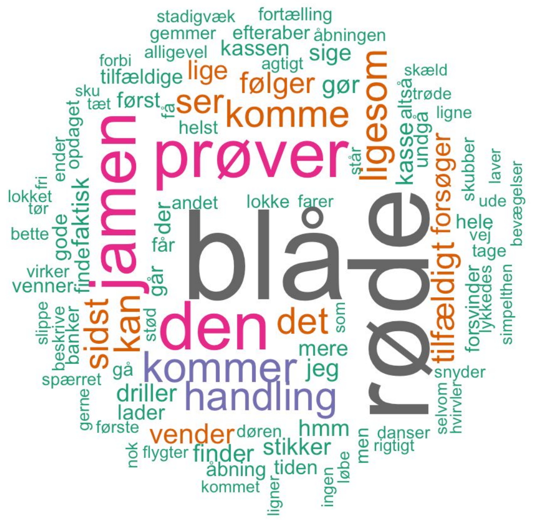
Figure A4: Term Frequency - Rest Term Frequency word clouds.

These words all appear more frequently in their class than in the other classes combined. The depressed used “prøver” about 1% of the time. It means “trying” and may refer both to themselves and the triangles. They also use the synonym “forsøger” (“attempting”).
While many of the highly scoring words in the depression class are in present tense, the metric finds mostly past-tense words for the schizophrenics.

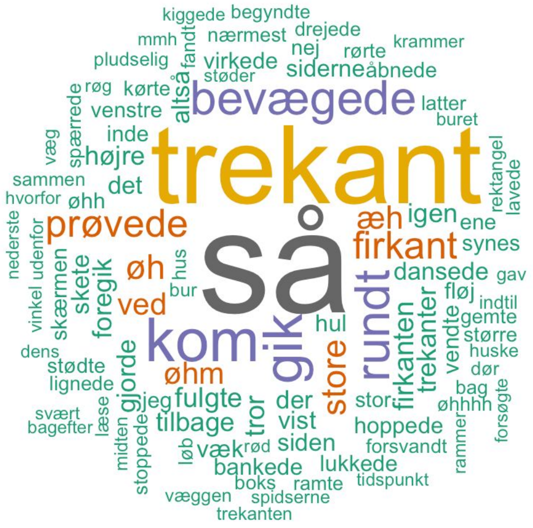
TF-NRTF (Term Frequency - Normalized Rest Term Frequency)



Control



Depression

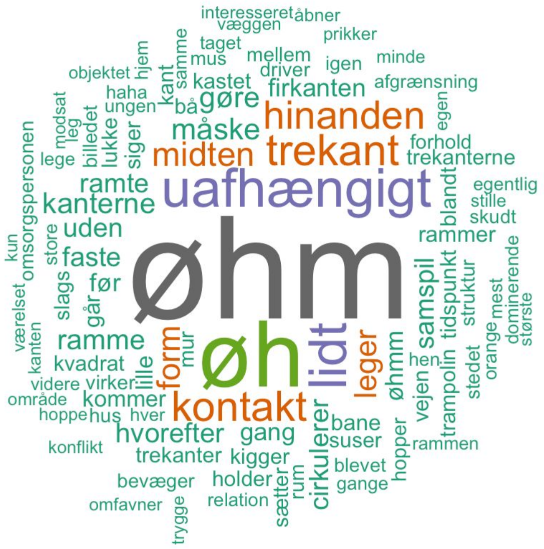


Schizophrenic

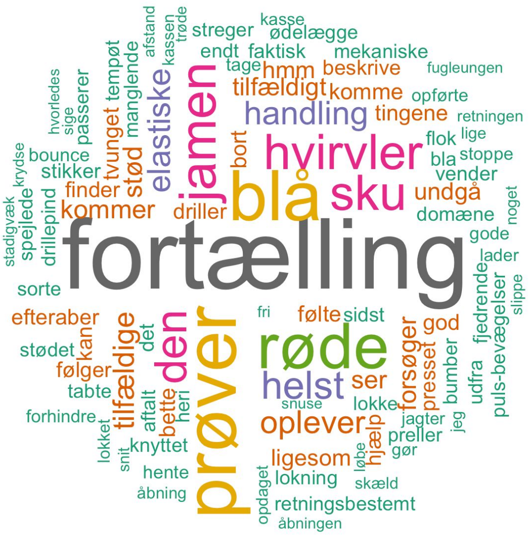
Figure A5: Term Frequency - Normalized Rest Term Frequency word clouds.

Here, we see some of the same words as in the term frequency clouds. “så” had a term frequency of 0.0965 among the schizophrenics with slightly less for the controls (0.0832) and depressed (0.0779). It could be argued that this relatively small difference is not enough to make the word interesting. In that case, we might want to use a different metric (like relative TF-NRTF), where the relative difference is also taken into account.
The word “blå” had a TF of 0.0414 among depressed, 0.0257 among controls and 0.0266 among schizophrenics, while “røde” had a TF of 0.0372 among depressed, 0.0234 among controls and 0.0238 among schizophrenics. Here, the relative differences are greater and it could be a relevant pattern, as these are both direct references to the two triangles.

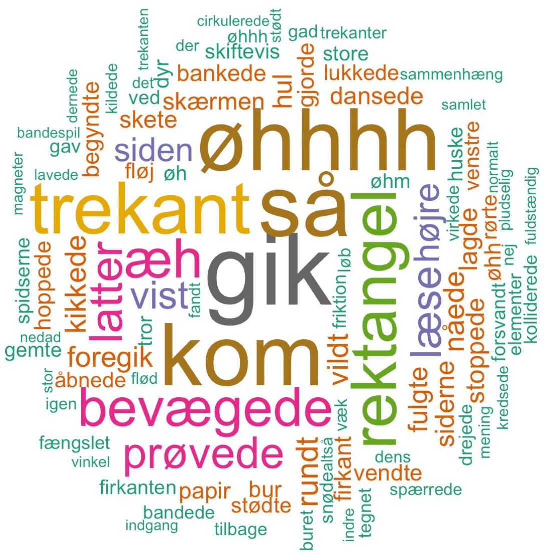
Rel TF-NRTF (Relative Term Frequency - Normalized Rest Term Frequency)



Control



Depression



Schizophrenic

Figure A6: Relative Term Frequency - Normalized Rest Term Frequency word clouds.

Some words are repeated from the TF-NRTF clouds, like “øhm”, “øh”, “så” and “gik”. The order has been changed though, and especially the word “fortælling” by the depressed has been deemed more distinguishing.

Appendix C: Highest scored words by metric

Word	Control Freq.	Depression Freq.	Schizophrenia Freq.	TF - RTF	TF - NRTF	Rel TF - NRTF	TF * IRF
uafhængigt	0.00050	0	0	0.00050	0.00050	0.00673	0.00035
faste	0.00029	0	0	0.00029	0.00029	0.00220	0.00020
samspil	0.00029	0	0	0.00029	0.00029	0.00220	0.00020
øhmm	0.00026	0	0	0.00026	0.00026	0.00185	0.00018
bå	0.00021	0	0	0.00021	0.00021	0.00124	0.00015
suser	0.00021	0	0	0.00021	0.00021	0.00124	0.00015
trampolin	0.00021	0	0	0.00021	0.00021	0.00124	0.00015
driver	0.00019	0	0	0.00019	0.00019	0.00098	0.00013
omsorgspersonen	0.00019	0	0	0.00019	0.00019	0.00098	0.00013
struktur	0.00019	0	0	0.00019	0.00019	0.00098	0.00013
fortælling	0	0.00076	0	0.00076	0.00076	0.03564	0.00053
hvirvler	0	0.00051	0	0.00051	0.00051	0.01584	0.00035
sku	0	0.00051	0	0.00051	0.00051	0.01584	0.00035
oplever	0	0.00038	0	0.00038	0.00038	0.00891	0.00026
aftalt	0	0.00025	0	0.00025	0.00025	0.00396	0.00018
bla	0	0.00025	0	0.00025	0.00025	0.00396	0.00018
bumber	0	0.00025	0	0.00025	0.00025	0.00396	0.00018
domæne	0	0.00025	0	0.00025	0.00025	0.00396	0.00018
drillepind	0	0.00025	0	0.00025	0.00025	0.00396	0.00018
endt	0	0.00025	0	0.00025	0.00025	0.00396	0.00018
øhhhh	0	0	0.00063	0.00063	0.00063	0.01967	0.00044
rektangel	0	0	0.00052	0.00052	0.00052	0.01366	0.00036
kikkede	0	0	0.00031	0.00031	0.00031	0.00492	0.00022
nåede	0	0	0.00031	0.00031	0.00031	0.00492	0.00022
vildt	0	0	0.00031	0.00031	0.00031	0.00492	0.00022
papir	0	0	0.00026	0.00026	0.00026	0.00342	0.00018
elementer	0	0	0.00021	0.00021	0.00021	0.00219	0.00015
fængslet	0	0	0.00021	0.00021	0.00021	0.00219	0.00015
friktion	0	0	0.00021	0.00021	0.00021	0.00219	0.00015
kolliderede	0	0	0.00021	0.00021	0.00021	0.00219	0.00015

Table A1: Ten highest scoring TF-IRF words for each diagnosis.

Word	Control Freq.	Depression Freq.	Schizophrenia Freq.	TF - RTF	TF - NRTF	Rel TF - NRTF	TF * IRF
øhm	0.01187	0.00025	0.00806	0.00356	0.00772	0.02189	-0.00481
kontakt	0.00107	0.00013	0.00021	0.00073	0.00090	0.00470	-0.00043
form	0.00131	0.00000	0.00073	0.00058	0.00094	0.00306	0.00000
uafhængigt	0.00050	0.00000	0.00000	0.00050	0.00050	0.00673	0.00035
leger	0.00167	0.00076	0.00042	0.00048	0.00107	0.00285	-0.00068
kanterne	0.00098	0.00013	0.00037	0.00048	0.00073	0.00250	-0.00040
hvorefter	0.00088	0.00013	0.00031	0.00044	0.00066	0.00225	-0.00036
gøre	0.00155	0.00038	0.00073	0.00043	0.00099	0.00258	-0.00063
cirkulerer	0.00057	0.00013	0.00005	0.00039	0.00048	0.00217	-0.00023
før	0.00112	0.00038	0.00042	0.00032	0.00072	0.00184	-0.00045
prøver	0.00385	0.01056	0.00230	0.00440	0.00748	0.02557	-0.00428
jamen	0.00699	0.01247	0.00366	0.00181	0.00714	0.01670	-0.00506
handling	0.00157	0.00649	0.00377	0.00115	0.00382	0.00924	-0.00263
forsøger	0.00143	0.00293	0.00042	0.00108	0.00200	0.00624	-0.00119
hmm	0.00062	0.00191	0.00031	0.00098	0.00144	0.00571	-0.00077
tilfældige	0.00017	0.00127	0.00016	0.00095	0.00111	0.00793	-0.00052
undgå	0.00012	0.00127	0.00031	0.00084	0.00106	0.00577	-0.00052
finder	0.00090	0.00242	0.00068	0.00083	0.00163	0.00486	-0.00098
driller	0.00126	0.00254	0.00047	0.00081	0.00168	0.00484	-0.00103
helst	0.00010	0.00089	0.00000	0.00080	0.00084	0.01175	0.00000
gik	0.00197	0.00064	0.00623	0.00362	0.00492	0.02314	-0.00253
kom	0.00343	0.00102	0.00785	0.00341	0.00563	0.01973	-0.00318
bevægede	0.00364	0.00089	0.00633	0.00180	0.00407	0.01128	-0.00257
æh	0.00098	0.00000	0.00267	0.00169	0.00218	0.01147	0.00000
prøvede	0.00202	0.00102	0.00450	0.00146	0.00298	0.00872	-0.00182
højre	0.00069	0.00013	0.00188	0.00107	0.00148	0.00649	-0.00076
vist	0.00026	0.00025	0.00157	0.00105	0.00131	0.00741	-0.00064
siden	0.00045	0.00013	0.00152	0.00094	0.00123	0.00602	-0.00062
foregik	0.00112	0.00013	0.00209	0.00085	0.00147	0.00479	-0.00085
siderne	0.00069	0.00013	0.00157	0.00075	0.00116	0.00425	-0.00064

Table A2: Ten highest scoring TF-RTF words for each diagnosis.

Word	Control Freq.	Depression Freq.	Schizophrenia Freq.	TF - RTF	TF - NRTF	Rel TF - NRTF	TF * IRF
øh	0.02186	0.00929	0.018	0.00000	0.00822	0.01322	-0.00887
øhm	0.01187	0.00025	0.00806	0.00356	0.00772	0.02189	-0.00481
lidt	0.01858	0.0151	0.0124	0.00000	0.00481	0.00652	-0.00753
trekant	0.03747	0.0240	0.0422	0.00000	0.00433	0.00497	-0.01519
hinanden	0.01397	0.0107	0.0107	0.00000	0.00326	0.00426	-0.00566
midten	0.00314	0.00064	0.00235	0.00015	0.00164	0.00337	-0.00127
lille	0.02746	0.0265	0.0254	0.00000	0.00151	0.00161	-0.01113
måske	0.00283	0.00076	0.00209	0.00000	0.00140	0.00271	-0.00115
firkanten	0.00490	0.00191	0.00513	0.00000	0.00138	0.00191	-0.00199
gang	0.00540	0.00344	0.00466	0.00000	0.00135	0.00179	-0.00219
blå	0.0257	0.04135	0.0266	0.00000	0.01516	0.02424	-0.01677
røde	0.0234	0.03715	0.0238	0.00000	0.01359	0.02167	-0.01506
prøver	0.00385	0.01056	0.0023	0.0044	0.00748	0.02557	-0.00428
den	0.00809	0.01578	0.00863	0.00000	0.00742	0.01402	-0.0064
jamen	0.00699	0.01247	0.00366	0.00181	0.00714	0.01670	-0.00506
kommer	0.00918	0.01107	0.00471	0.00000	0.00412	0.00658	-0.00449
handling	0.00157	0.00649	0.00377	0.00115	0.00382	0.00924	-0.00263
ligesom	0.00749	0.01031	0.00607	0.00000	0.00352	0.00536	-0.00418
komme	0.00776	0.01056	0.00644	0.00000	0.00346	0.00516	-0.00428
det	0.00971	0.01501	0.0137	0.00000	0.00333	0.00430	-0.00609
så	0.0832	0.0779	0.09650	0.00000	0.01594	0.01985	-0.03913
trekant	0.0375	0.0240	0.04223	0.00000	0.01147	0.01598	-0.01712
kom	0.00343	0.00102	0.00785	0.00341	0.00563	0.01973	-0.00318
gik	0.00197	0.00064	0.00623	0.00362	0.00492	0.02314	-0.00253
rundt	0.0172	0.0137	0.01962	0.00000	0.00418	0.00534	-0.00796
bevægede	0.00364	0.00089	0.00633	0.0018	0.00407	0.01128	-0.00257
prøvede	0.00202	0.00102	0.00450	0.00146	0.00298	0.00872	-0.00182
firkant	0.0129	0.0126	0.01565	0.00000	0.00291	0.00360	-0.00634
øh	0.0219	0.00929	0.01800	0.00000	0.00243	0.00282	-0.0073
store	0.0200	0.0174	0.02114	0.00000	0.00241	0.00274	-0.00857

Table A3: Ten highest scoring TF-NRTF words for each diagnosis.

Word	Control Freq.	Depression Freq.	Schizophrenia Freq.	TF - RTF	TF - NRTF	Rel TF - NRTF	TF * IRF
øhm	0.01187	0.00025	0.00806	0.00356	0.00772	0.02189	-0.00481
øh	0.02186	0.00929	0.018	0.00000	0.00822	0.01322	-0.00887
uafhængigt	0.00050	0	0	0.00050	0.00050	0.00673	0.00035
lidt	0.01858	0.0151	0.0124	0.00000	0.00481	0.00652	-0.00753
trekant	0.03747	0.0240	0.0422	0.00000	0.00433	0.00497	-0.01519
kontakt	0.00107	0.000130	0.00021	0.00073	0.00090	0.00470	-0.00043
hinanden	0.01397	0.0107	0.0107	0.00000	0.00326	0.00426	-0.00566
midten	0.00314	0.00064	0.00235	0.00015	0.00164	0.00337	-0.00127
form	0.00131	0	0.00073	0.00058	0.00094	0.00306	0.00000
leger	0.00167	0.00076	0.00042	0.00048	0.00107	0.00285	-0.00068
fortælling	0	0.00076	0	0.00076	0.00076	0.03564	0.00053
prøver	0.00385	0.01056	0.0023	0.00440	0.00748	0.02557	-0.00428
blå	0.0257	0.04135	0.0266	0.00000	0.01516	0.02424	-0.01677
røde	0.0234	0.03715	0.0238	0.00000	0.01359	0.02167	-0.01506
jamen	0.00699	0.01247	0.00366	0.00181	0.00714	0.01670	-0.00506
hvirvler	0	0.00051	0	0.00051	0.00051	0.01584	0.00035
sku	0	0.00051	0	0.00051	0.00051	0.01584	0.00035
den	0.00809	0.01578	0.00863	0.00000	0.00742	0.01402	-0.00640
helst	0.0001	0.00089	0	0.00080	0.00084	0.01175	0.00000
handling	0.00157	0.00649	0.00377	0.00115	0.00382	0.00924	-0.00263
gik	0.00197	0.00064	0.00623	0.00362	0.00492	0.02314	-0.00253
så	0.0832	0.0779	0.09650	0.00000	0.01594	0.01985	-0.03913
kom	0.00343	0.00102	0.00785	0.00341	0.00563	0.01973	-0.00318
øhhhh	0	0	0.00063	0.00063	0.00063	0.01967	0.00044
trekant	0.0375	0.0240	0.04223	0.00000	0.01147	0.01598	-0.01712
rektangel	0	0	0.00052	0.00052	0.00052	0.01366	0.00036
æh	0.00098	0	0.00267	0.00169	0.00218	0.01147	0.00000
bevægede	0.00364	0.00089	0.00633	0.00180	0.00407	0.01128	-0.00257
latter	0.000070	0	0.00073	0.00066	0.00070	0.00916	0.00000
prøvede	0.00202	0.00102	0.00450	0.00146	0.00298	0.00872	-0.00182

Table A4: Ten highest scoring Relative TF-NRTF words for each diagnosis.
For the word “trekant”, we see that it’s actually more frequent with the schizophrenics than the controls, but included in both sets. Here, the Relative TF-MRTF (M=Maximum) could possibly perform better.

Appendix D: Results for upsampled dataset

On the upsampled dataset, the Bernoulli Naïve Bayes model achieved the highest balanced accuracy of 0.652 (baseline = 0.5). With an overall accuracy of 0.537 (baseline = 0.334) and an F1 score of 0.538 (baseline = 0.334), it performs well above chance. Table A5 shows the results for each of the model architectures.

Model	**N Grams**	**Overall Accuracy**	**Balanced Accuracy**	**F1**	**Precision**	**Recall**
BSL Mean (SD)		0.334 (0.00369)	0.500 (0.00276)	0.334 (0.00369)	0.334 (0.00369)	0.334 (0.00369)
BSL Max.		0.343	0.507	0.343	0.343	0.343
SVM	1	0.481	0.611	0.480	0.500	0.481
SVM SW	1	0.457	0.593	0.451	0.485	0.457
NB Multinomial	1	0.514	0.636	0.516	0.528	0.514
NB Multinomial SW	1	0.490	0.618	0.492	0.502	0.490
NB Bernoulli	1	0.537	**0.652**	0.538	0.542	0.537
NB Bernoulli SW	1	0.517	0.638	0.518	0.522	0.517
NB Compl.	1	0.498	0.623	0.499	0.501	0.498
NB Compl. SW	1	0.484	0.613	0.483	0.484	0.484

Table A5: Best results for each of the model architectures on the upsampled dataset, with and without subwords (SW). The *BSL* rows are the baseline evaluations.

Figure A7 shows the confusion matrix. 54% of the controls are correctly identified, with 29.5% predicted to be schizophrenic and 16.5% to be depressed. 56.7% of the depressed patients are correctly identified, with 25% of them predicted to be controls and 18.4% of them schizophrenic. 50.3% of the schizophrenics are correctly identified, with 33.8 being predicted to be controls and 15.9% depressed.

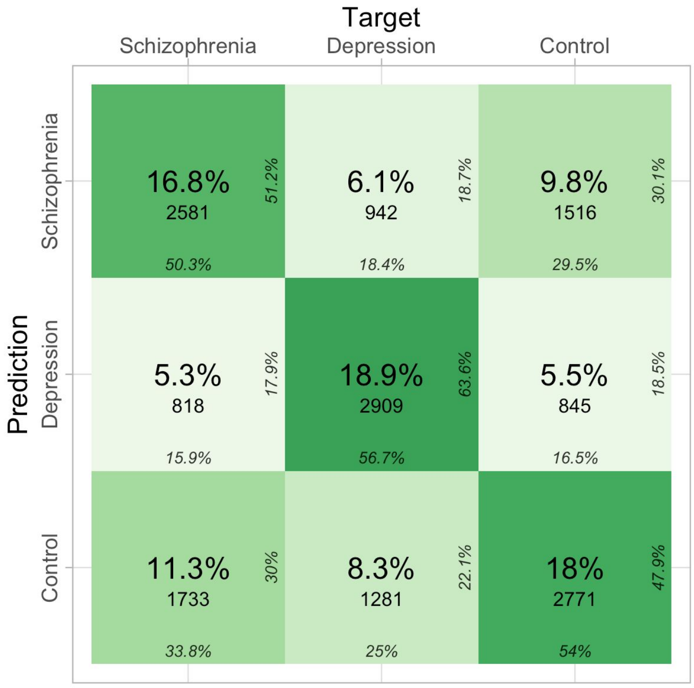


Figure A7: Confusion matrix for the best model on the upsampled dataset.