Universiteit Leiden
The Netherlands

Opleiding Informatica

Depression prediction using twitter data:

Comparing LIWC and Bag-of-Words features

C. S. Heath

Supervisors:
Suzan Verberne & Anne Dirkson

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

24/07/2020

Abstract

Depression is a mental health issue that is still being underdiagnosed to this day. This is the main reason that the idea of being able to detect depression via social media data is explored. Social media is widely used by people to gain information and share thoughts and feelings. In this research we aimed to compare models that can detect if a user might possibly have signs of depression based on a subset of their tweets. We tried to do so by awnsering the question: to what extent is it possible to detect if a person might have depression signs based on Twitter text data using classification models? In the process of doing so, two different feature sets were compared, one made by the Linguistic Inquiry and Word Count tool and the other a Bag of Words feature set that uses words as features. For each feature set three classifiers were compared: Naive Bayes, Random Forest and Linear Support Vector Machine (SVM). The best performing classifier on the LIWC feature set after hypertuning the parameters was Random Forest which scored an accuracy of 82%. For the BOW feature sets both the data set including the tweet on which users were labeled as depressed, the signal tweet, as a modified version which did not include the feature test were compared: the one including the signal tweet had an accuracy of 94% using Linear SVM and 71% without the signal tweet using Linear SVM. We found that while the accuracy score seemed promising for the BOW models, the LIWC feature models gave us more useful insight and correspond the best with previous research.

Contents

1 Introduction

Social media has greatly impacted the amount of data that is produced every day on the internet. According to a report by DOMO, a mobile, cloud-based operating system service, every minute worldwide there are over 92,340 posts on Tumblr, 55,140 new photos on Instagram and 511,200 tweets posted on Twitter [1]. Besides communicating with other people, users use social media to express their feelings, emotions and thoughts via text or images [2]. While there has been a lot of research on how social media can impact the mental health of someone [3][10], at the same time researchers are also trying to find ways to predict if someone struggles with mental health issues via the behavior they show on social media [18][20]. The majority of these studies focus on depression.

1.1 Online expression of depression

While many people feel sad or depressed from time to time, those feelings often go away after a few days. However, when these feelings do not go away after a few days, someone might suffer from depression. According to the World Health Organization [21] the most common signs of depression are a low or sad mood, lack of appetite, loss of interest and enjoyment, and reduced energy levels. These signs are often combined with common anxiety signs such as having a low self-esteem and feeling irritable and anxious. People who are suffering from anxiety can also experience suicidal thoughts or thoughts of harming themselves. Worldwide it is estimated that around 264 million people of all ages suffer from depression [21]. Even though there is a lot more attention for this condition, especially adolescent depression continues to being under diagnosed by primary care physicians [17].

Because social media provides a platform for a lot of people to share their thoughts,feelings and opinions, it is a very interesting source of data for researchers who are interested in human behavior combined with online expression of emotions and feelings. In their study, Ophir et al. compared the traditional offline clinical view of depression with the online expression via Facebook [13]. To measure the offline clinical expression, they used the Depression DSM-5 Diagnostic criteria, which states that an individual must be experiencing at least five of the given symptoms during a consecutive 2-week period. At least one of those symptoms should be a depressed mood or loss of interest and enjoyment. They found that depressive status updates were more likely to include cognitive distortions such as black and white thinking, selective thinking and over generalization, more lyrical, abstract in concept, and also had more negative attitudes towards others. Their findings suggest that online expression of depression usually involve references to the more psychological and behavioral aspects of depression, while the offline image of depression is mostly focused on more physical symptoms such as described in the paragraph above, such as being tired, lack of appetite.

The way in which users use social media can also say something about their condition, as Park et al found [14]. They found that non-depressed individuals mostly used Twitter as a source of information consuming and sharing tool, while depressed users used Twitter more for social awareness, emotional interaction, to console themselves and express how they were feeling. Non depressed users were more keen to manege the type and amount of information they saw on Twitter, rather than make sure to control the type of sentiments they received. This difference was measured

by asking people the reason for following people on Twitter. One of the reason for this, according to this study, might be the mental state of a person could influence the perception of the world and the people around them. So depressed people look at the world in a different light than people who are not depressed. Which makes sense if you look at the symptoms given by the WHO and DSM-5, they often include loss of enjoyment and interest, but also a low or sad mood.

1.2 Research questions

Because depression is still being under diagnosed, this research will attempt to investigate if it is possible to recognize signs of depression in social media messages. This will be done by answering the following question:

"To what extent is it possible to detect if a person might have depression signs based on Twitter text data using classification models?"

To answer this question this study will also compare two widely used feature sets: using the Linguistic Inquiry and Word Count tool to generate features and the Bag of Words model features. The Linguistic Inquiry and Word Count tool [16], or LIWC for short, is a text analysis application which can be used to analyze text based on word categories. The word categories are composed by a group of psychology researches and focus on the various emotional, cognitive, and structural components of someone's written speech.
We will try to compare these two feature sets by answering the following sub question:

"How do the results of classifying with LIWC features compare to classifier results with the Bag of Words model?"

1.3 Thesis overview

In this thesis the extent to which a classifier model can predict depression based on tweets will be explored, as well as the comparison between two different feature sets: LIWC and bag of words. The background context as well as an explanation of LIWC will be given in section 2; section 3 describes the data collection, sampling and preprocessing of the data as well as the classification process; section 4 describes the results of the classification models and the analysis of the results. The discussion takes place in section 5 and finally the conclusions and suggestion for further research will be given in section 6.

2 Background

The following subsections contain an overview of the studies related to this thesis and background information on the software that is used in this research: LIWC.

2.1 LIWC

For part of the comparison in this research we use the Linguistic Inquiry and Word Count tool, LIWC for short, to generate the percentages of words that indicate different emotions, thinking styles, but also linguistic categories such as tenses, used in a sentence [16]. LIWC was created by researches who had an interest in psychology who managed to develop categories that capture the social and psychological state of people through their language. So what LIWC does is it counts per line of data that you give it to analyze, the words that reflect the different categories. These word counts are then converted to the percentage of that word relative to the whole line. Words can be part of multiple categories, an example taken from the LIWC manual itself is that the word 'cried' is part of five different word categories: Sadness, Negative Emotion, Overall Affect, Verb, and Past Focus.

LIWC has approximately 90 categories which is returns as output. These categories are summarized by:

- Summary Language Variables, which include some general information like words per sentence, emotional tone and number of words that are also present int the dictionary.
- Linguistic Dimensions, like pronouns, prepositions, common adverbs and negations, which focuses on how we communicate.
- Other Grammar, such as common verbs, adjectives, numbers and comparisons.
- Psychological Processes, which is the biggest category including all kinds of psychological processes and social interactions like positive or negative emotion, anxiety, sadness, family, health, time orientations and informal languages such as swear words and netspeak (words like btw, thx and lol)

For all 90 categories a separate list or dictionary was built consisting of relevant words. After creating a master list of all the words the researchers could think of that reflected that category, a team of human judges evaluated each word and checked if that word truly reflected the category. The words is only included in the final dictionary if all the judges decided that is was appropriate.

It is important to note that LIWC does not take context into account. So a person might use a lot of depression related words, but would that mean they are really depressed? It is very difficult to prove as one could, for example, be pretending to be depressed. This also mean that LIWC does not take figure of speech such as sarcasm, metaphors or irony into account. So we should keep this in mind when analyzing the results given by LIWC.

2.2 Related work

Over the past decade there has been a lot of research to predict mental health issues using social media data. Islam et al. showed a machine learning approach to predict depression in Facebook users [9]. They focused on four types of factors like emotional process, linguistic style, temporal process and all three combined. To find these factors they also used LIWC to get values of each type of factor. After comparing four major classifiers: Support Vector Machine, K-Nearest Neighbors, Decision trees and Ensemble, they found that Decision trees performed best at predicting depression. While all classifiers performed between 60 and 80%, the decision trees gave the highest score for recall and F-measure for the depressed class.
Similar research was done predicting postpartum depression using machine learning techniques on Reddit data. They found strong predictive capabilities with an accuracy of 87% for postpartum depression and an accuracy of 91.7% for predicting depressive content using multiplayer perceptron neural network.

LIWC has been a popular tool for not only predicting depression and other mental health issues, but also to give insight in the ways in which depressed people interact and represent themselves on online platforms [12]. Besides the frequently used LIWC software, the bag of words features are also widely used within research in depression classification. In a extensive evaluation of the effectiveness of using social media activities to estimate depression levels, it is shown that the highest accuracy for a bag of words feature set is somewhere around 69% [20]. In this research in addition to the bag of words feature set based on Japanese tweets, they also used features like topic, positive and negative words, time and tweet frequency to predict depression levels. A accuracy of 61% was found when only using bag of words with a total of 20.000 words with a appearance of at least 25 across all users, and 62% when using 2000 words. They mention overfitting as a cause for the low accuracy result.

An accuracy of around 70% seems to be a good average for most classification models for prediction depression levels. One of the most cited researches within this fields also found an accuracy of 70% with an SVM classifier [7]. Besides LIWC, this research also focused on what people talk about instead of just the linguistic style of depressive language. To achieve this a depression lexicon was created by extracting words from the "Mental Health" category of Yahoo! Answers.

There have also been a few shared tasks that were centered around detection of (mental) health problems. The CLEF eRisk labs have been centered around the early risk prediction for different purposes. The eRisk of 2020 was focused on the early detection of sign of self-harm and measuring the severity of the signs of depression [11]. In 2015 the CLPsych shared tasks was centered around depression and PTSD on Twitter [5]. Within these tasks they aimed to compare various approaches to modeling language relevant to mental health (Depression and PTSD specifically) from Twitter. They found that there were clear benefits to using topic modeling approaches as these provided strong signals relevant to mental health, but also that simple linguistic features provided some classification power.

This thesis uses a data set that was constructed by Shen et al [18]. In their research they aimed to create a multi model learning solution because according to their paper the behavior of social media users are multi-dimensional. The multimodal depressive dictionary learning model they propose

combines different modalities that are not independent of each other and share common patterns that cannot be captured otherwise. They extracted from each user in their database depression related features, that were not only inspired by the offline symptoms of depression but also by social media behavior. They divided these features in six modalities: social network features, user profile features, visual features, emotional features, topic-level features, and domain-specific features. For the emotional features they used among other things LIWC to extract the amount of positive and negative emotion words. Their final model considers, besides each feature group as a single modality, the relations between all modalities. No report was given on the relevance of these features. They did mention some interesting findings related to the features they used, such as that depressed people tend to post more between 23:00 and 6:00, and that depressed used use more emotional words in their tweets.
They compared four different classification methods: Naive Bayes, Multiple Social Networking Learning, Wasserstein Dictionary Learning and their own Multi modal Depressive Dictionary Learning. They found that Naive Bayes achieved a accuracy of 73%, a recall of 73% and an F1 score of 72%. Their own Multimodal Dictionary Learning achieved the best performance with a 85% F1-score.

3 Methods

The goal of this research is to try and predict whether a user might have signs of depression based on their tweets, and in this process comparing two different feature sets. In this section the steps to perform this prediction are explained. All the steps described below relating to the data preparation, sampling and preprocessing are written in Python, using mainly the Scikit-learn module [15].

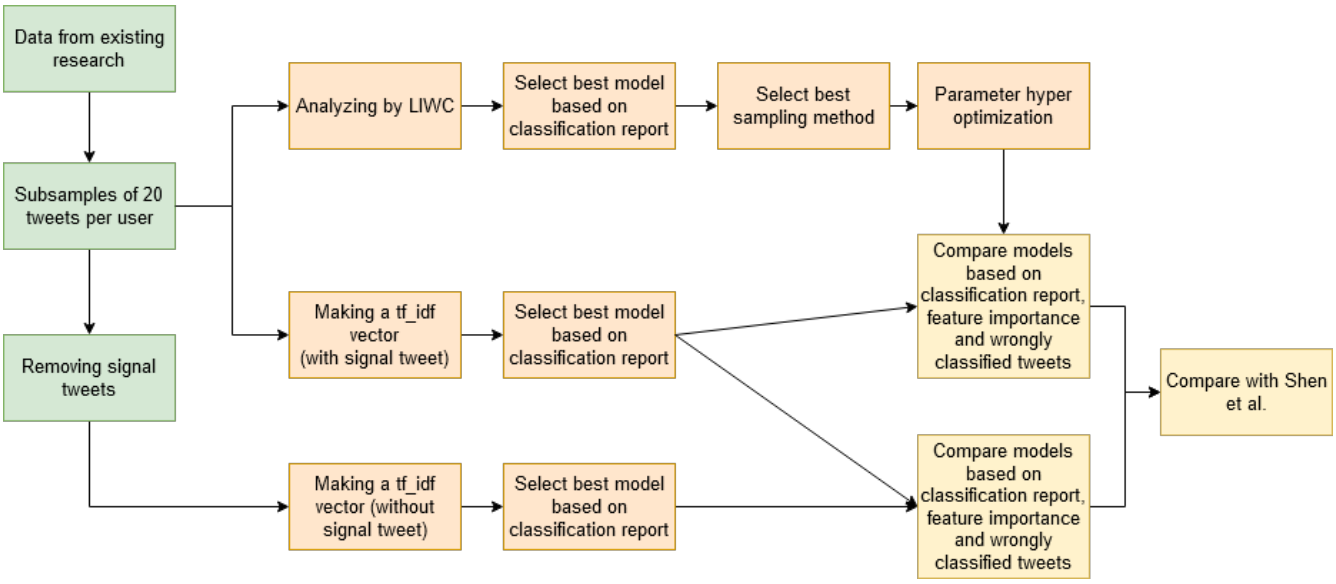

Figure 1: Workflow diagram

Figure 1 shows an illustration of the method that will be used in this research. The green squares show preprocessing the data, the orange squares show the classification steps and the yellow squares show the results that will be discussed in section 4.

3.1 Data collection

The data that is used in this research is collected by Shen et al [19] for their multi modal dictionary learning solution mode to detect depression in social media [18]. The data set they provided consisted of three different sub data sets:

Data set D1 contains tweets from users who had their anchor tweet, which is their most recent tweet or the tweet a users highlights on their profile, between 2009 and 2016 satisfying the strict pattern "(I'm/ I was/ I am/ I've been) diagnosed depression". These users were labeled as depressed and the tweet on which this was based is from here on indicated as the signal tweet. All other tweets tweeted within one month from the anchor tweet were also included. However, after going through the data set during the research, we found that this pattern was not at all as strict as they claimed. There were a lot of cases where the anchor tweet did not have the string as one contiguous string, which seemed implied in their introduction. For example the following tweets were used to classify someone as depressed: "The past few weeks I was diagnosed with Anxiety and Depression ", " I was diagnosed today with very serious anxiety and depression", " I was diagnosed again for the

7th time with depression and anorexia for 3rd time". The pattern is there but it not as strict as they claimed in their data set description. But because the pattern was not strict they were able to collect more data. They did seem to check that there were no negations in the tweets that would mean that the users was not diagnosed with depression because there were no such tweets to be found.

Data set D2 was constructed by selecting users who had never posted any tweet containing the string "depress". These users were labeled not depressed. And again all other tweets tweeted by these users one month from their anchor tweet were selected.

Data set D3 is an unlabeled data set for depression-candidates. This means that users were obtained if their anchor tweets loosely contained the string " depress".
Because we have no interest in an unlabeled data set for this research, data set D3 is ignored. Only data set D1 and D2 are used in this research. The paper from this research states the following statistics on their data:

Table 1: Data set statistics according to the paper

Data set	D1	D2	D3
Users	1,402	> 300 million	36,993
Tweets	13,130	> 10 billion	25,076.677

However, in the data set they delivered there were a lot more user timelines labeled positive: it contained 2.626 timelines instead of the 1.402 they described in their paper. This is likely due to a lot of people tweeting more than could fit in one file, so the timelines are split up for some users. We treated every timeline as a different user, so some users appear more than once. Another difference is the amount of negative timelines we were able to download. It is never stated by Shen et al that all data is given in the data set they provided on their website. We did try and download the data set multiple times to see if that might have been the problem, but we got the same result every time. However, we do not foresee this to be a problem because 2,626 positive combined with over 300 million negative timelines would cause a hugely imbalanced database.

3.2 Data preparation

3.2.1 Sampling

After selecting a smaller data set from the one published by Shen et al, we decided to take the same approach as them by not classifying individual tweets but classifying users. This is done to achieve more data to learn from. Doing this will probably be more realistic than to classify just one tweet containing a maximum of 140 characters (between 2009 and 2016 Twitter did have a character limit set on 140, since 2017 this is doubled to 280). However, not every user has the same amount of tweets in their timeline. For the 7999 users in our database the person who had the least tweets had a total of two tweets in their database entry while the person with the most tweets had 5337 tweets, but as we can see in Figure 2 this is an outlier in the data. The mean number of tweets per user in the data set was 595 while the median is 228. This shows that there is a large deviation between the amount of tweets per user in the data set.

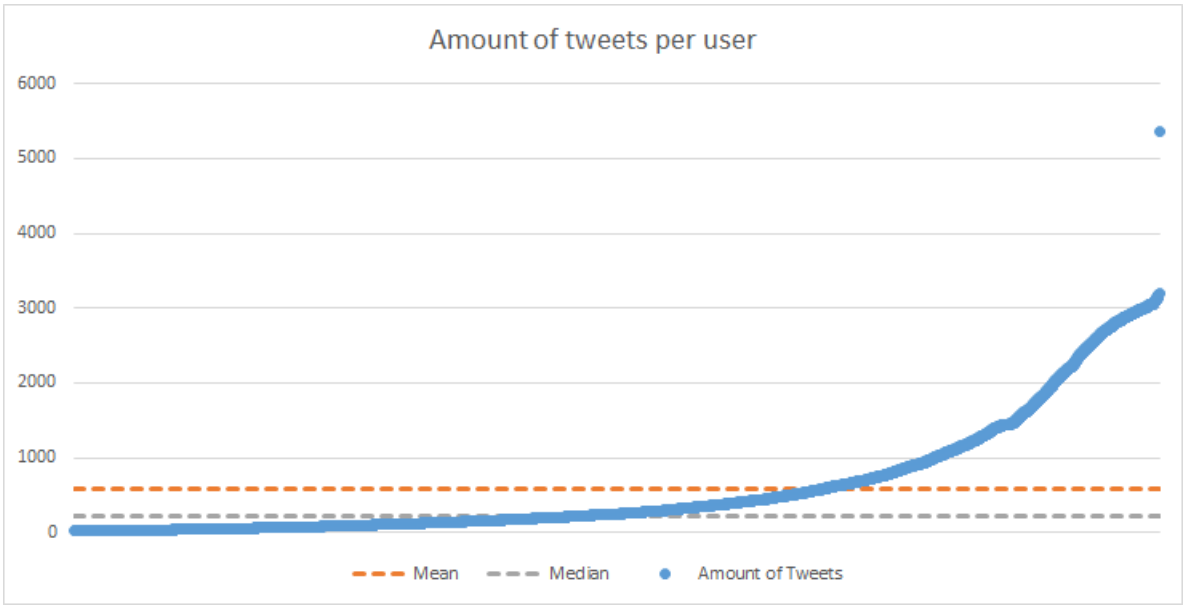

Figure 2: Tweet count per user

Because of the large deviation in number of tweets, we decided to bundle some tweets per user so that every item has the same amount of data. This is done by taking 5 random samples of 20 tweets per user with replacement. This means that for users who only tweeted 20 tweets, there will be 5 data sets which each can contain duplicate tweets because of replacement sampling. This makes the chance of identical subsets a lot smaller. This also prevents the imbalance between the different data points, so data points that would otherwise contain 1000 tweets or only 50 tweets will after sampling both be distributed over 5 random samples of 20 tweets. This resulted in the data set with the statistics shown in Table 2.

Table 2: Data set after sampling, where positive means labled as depressed and negative means labeled as not depressed

	Positive	Negative
User Timelines	2,626	5,373
Random subsets	13,130	26,830

3.2.2 Data preprocessing

The data contained a number of non-English tweets. Probably because of the method used to create the D2 data set. In this research we are only interested in the English tweets, as the LIWC library we used was written in English. To remove most of these tweets that were not written in English, the Lexical normalization pipeline [6] was used[1]. By removing most of the non-English tweets, we removed 0.017% of the total amount of tweets, by removing the spelling error, 0.0002% of the words changed. The pipeline also removes some spelling errors in the tweets. We decided to do this because we wanted to focus on the kind of words people use in their tweets.

[1] https://github.com/AnneDirkson/LexNorm

LIWC does not recognize words with spelling mistakes as the correct words. We later found that the removal of spelling mistakes did not impact the accuracy and positive F1 score of the models a lot.

To generate the features, LIWC was used to analyze and categorize the data and return a vector with the results. To generate a Bag of Words feature vector the TfidfVectorizer function was used. With the Scikit-Learn package in Python we created a term frequency–inverse document frequency (or TF-IDF for short) matrix which stores the term weight of the words per data item. The term frequency is the number of times a term occurs in the document (in this case in the data set of tweets). To create the tf-idf term weighing, the term frequency is multiplied by the idf component: $tf\text{-}\ idf(t,d) = tf(t,d) \times idf(t)$. Where t is the term and d the document. The idf component is defined as follows:

$$idf(t) = log\frac{1+n}{1+df(t)} + 1,$$

where n is the total number of documents in the document set (tweet subsamples in the whole data set) and $df(t)$ is the number of subsets that contain the term t.

Because this approach could not be executed on the computer we used without any adjustments, we changed the minimum document frequency from the default value 1 to 25 and 50, so that all features would appear at least 25 and 50 times in the combined corpus. The minimum document frequency (min_df) is a parameter from which its value decides how many instances in the whole set need to contain that word to end up in the word vector. We tried both 25 and 50 because this way we could eliminate words that would occur to little for two reasons: first to study the effect of the dimension and word count of the vector and second to prevent the classifier from over fitting to words that are to specific.

For the first comparison between the LIWC and BOW approaches we kept the signal tweets in the data set because Shen et al. didn't mention that they removed the signal tweets before training their model. However, as shown in section 4, it seemed like the BOW approach was quite biased towards some specific words, because the features that had the biggest weight were also present in the pattern on which the signal tweets were selected. So for a second comparison we did remove all the signal tweets from the data set to see if that made a difference.

3.3 Models

The classifiers that we compared are the three most used classifiers in the text analysis field: Naive Bayes, Linear Support Vector Machine and Random Forests. Our data had a ratio of one user that was labeled as positive for every two users that were labeled as negative. Because of this small imbalance we decided to include under- and oversampling in the comparison between classifiers for each approach. The three kinds of sampling that we test are: Random Oversampling, Random Undersampling and Synthetic Minority Over-sampling Technique (SMOTE) [4]. Random Oversampling is achieved by randomly duplication examples in the minority class while and Random Undersampling is achieved by randomly deleting examples in the majority class. Which means oversampling gives a bigger data set while undersampling gives a smaller dataset. SMOTE is a technique that applies oversampling but by synthesizing new examples to provide additional information. This is done by selecting samples that are close in the feature space drawing a line

between them and drawing a new sample at a point along that line.
For each classifier we train the model without any of these sampling methods. We take the sampling methods into consideration for the one that performed best for each approach. We ended up only taking sampling into account for the LIWC features, as we discovered that the our tf-idf feature vector had issues we wanted to explore first before applying any form of over or under sampling. It was not really useful to apply sampling with those results.

After training our models and comparing them on their standard parameters without sampling, we optimized the parameters for the best performing model. This ended up being the Random Forest classifier as seen in section 4. The set of hyper optimized parameters used for the final Random Forests model was created on the train set with the help of the GridSearchCV function from Scikit-learn, which executes an exhaustive search over specified parameter values for the model. The following grid was used to find the optimal values: n_estimators from 200 to 2000 with intervals of 100, for max_features only two parameters are possible: 'auto' and 'sqrt', for max_depth a range of 10 to 110 with intervals of 10 and for bootstrap 'True' and 'False'. The bootstrap parameter indicates if bootstrap samples are used when building trees. If the parameter is set at 'False' the whole data set is used to build each tree. The parameter that differ from the standard parameters are shown in Table 3.

Table 3: Results hyperparameter optimization

Parameter	**Value**
n_estimators	1400
max_features	auto
max_depth	40
bootstrap	False

For the Bag of Words features the same approach was used to compare all three models with their standard parameters. The Linear Support Vector Machine and Random Forest were both promising so we continued with those two. But because of the results of the Scikit-learn classification report for both, we did not get around to the hyper optimization of the parameters. We decided not to do this because we found that the LIWC features gave us a better model for predicting signs of depression and decided to continue optimizing that one.

3.4 Evaluation

A train/test split of 70/30 is created by using the train_test_split function in Scikit-learn to reduce the chance of over fitting. A hyperparameter optimization is applied via cross validation on the train set to get the best performing parameters as seen in the paragraph above.

The evaluation measures that will be considered the most important to compare the different classification models are the overall accuracy, positive recall, precision and positive F1 score. The decision to choose only the positive recall, precision and F1 score besides accuracy was made because those were of most interest to this research. Recall is calculated by the following formula:

$$\frac{TruePositive}{TruePositive + FalseNegative}$$

For the positive Recall goes that in this case the true positives are users that were correctly classified as 'True' and the false negatives are users that were classified as 'False' while they had the label 'True'.
The positive Precision is calculated by the following formula:

$$\frac{TruePositive}{TruePositive + FalsePositive}$$

The F1 score shows a mean of both precision and recall. Only the positive factors are shown here because we are most interested in the people who might be showing signs of depression instead those who are not.
The feature importance and wrongly classified users will also be discussed in the results section. These subjects will be investigated to get more insight into what the models actually suggest and why they classify certain subsets in a certain way.

To determine which model is the best we will be looking at which model gives the most useful prediction. This is done by taking into account the most important features, the accuracy, recall and F1 score, and the users that were wrongly classified.

4 Results

In this section we will discuss the results obtained by this research. First we will address the results of the model quality and feature importance for both the LIWC and the BOW features. After that we will discuss the results of the wrongly classified users.

4.1 Model performance

As discussed in section 3 we first tried several models for the LIWC features without using any form of over or under sampling. Using the Random Forest an accuracy of 0.82 was achieved, while using Linear Support Vector machine ended up with an accuracy of 0.71 and Naive Bayes came in the lowest with an accuracy of 0.61. Based on the accuracy of the models as seen in Table 4, one of the three was chosen to explore further. As explained in section 3 we chose to continue with the Random Forest as model for the LIWC features.
Table 5 shows the results of the comparison between the different sampling methods after optimizing the Random Forest model without any form of sampling.

Table 4: Comparing different classifiers for LIWC features

	Random Forest	Linear Support Vector Machine	Naive Bayes
Accuracy	**0.823**	0.712	0.610
Positive Recall	0.612	0.514	**0.814**
Positive Precision	0.814	**0.832**	0.449
Positive F1 score	**0.708**	0.628	0.579

Table 5: Comparing different sampling methods on Random Forest with LIWC features

	No Sampling	Random Undersampling	Random Oversampling	SMOTE
Accuracy	**0.823**	0.796	**0.823**	0.820
Positive Recall	0.612	**0.838**	0.584	0.705
Positive Precision	0.814	0.648	**0.830**	0.738
Positive F1 score	0.708	**0.731**	0.686	0.721

These results show that sampling did have an increasing effect in some of the cases. We can see in Table 5 that the random oversampling and SMOTE methods showed an higher accuracy while random undersampling achieved the highest positive recall. The F1 score for all three methods is the same. We decided to go for the SMOTE approach because based on the numbers in Table 5 is performs better than random oversampling but ends up with a larger data set than if we were to use random undersampling.

Next, we will take a look at the results for the Bag of Words features. The same method was applied to the BOW features. The most promising methods were Linear Support vector machine with an accuracy of about 0.94 and Random Forest with an accuracy of about 0.92. We did not bother to investigate Naive Bayes any further because this model had an accuracy of 0.70 as seen

in Table 6. For BOW we made a distinction between a minimum document frequency (min_df) in words. The results of the models are shown in Table 7.

Table 6: Comparing different classifiers for BOW features

	Random Forest	Linear Support Vector Machine	Naive Bayes
Accuracy	0.923	**0.943**	0.696
Positive Recall	0.779	0.868	**0.939**
Positive Precision	0.984	0.955	0.452
Positive F1 score	0.869	**0.959**	0.670

Table 7: Comparing different values of the min_df parameter

	Min_df = 25 Linear SVM	Min_df = 50 Linear SVM	Min_df = 25 RF	Min_df = 50 RF
Accuracy	**0.943**	0.938	0.923	0.927
Positive Recall	**0.868**	0.860	0.779	0.791
Positive Precision	0.955	0.948	**0.984**	0.982
Positive F1 score	**0.959**	0.902	0.869	0.876

The results in the table show that there is not much of a difference between using 25 or 50 as a min_df of the feature words. 25 was the lowest we could go to be able to compute the vector in a workable time frame. Using a minimum document frequency of 25 gave a vector of 12178 word features, while using a minimum document frequency of 50 gave 7229 features. Because there was not much of a difference, only the figures for a minimum document frequency of 25 will be shown, the other figures will be included in the appendix. The very high accuracy results were not in line with what previous research that used the Bag of Words features for prediction depression showed.

After we inspected the models (most important features), we decided to remove the signal tweets from the data and re-run the experiments. The feature importance in Figure 5 and 4 showed that certain words on which the signal pattern was based had a very high importance. While in the paper written by Shen et al [18] it is never stated that they did not include the signal tweets, we did try and remove those signal tweets and let the models run again. By removing the signal tweets from the dataset the the results shown in Table 8 were obtained.

Table 8: Comparing different values of the min_df parameter without the signal string

	Min_df = 25 Linear SVM	Min_df = 50 Linear SVM	Min_df = 25 RF	Min_df = 50 RF
Accuracy	**0.713**	0.695	0.710	0.712
Positive Recall	**0.418**	0.377	0.138	0.145
Positive Precision	0.596	0.559	**0.925**	0.912
Positive F1 score	**0.491**	0.450	0.240	0.250

Again both 25 and 50 were used as minimum document frequency. This time using a minimum document frequency gave a vector of 12563 word features, while using a minimum document

frequency of 50 gave 7478 features. These accuracy scores seem more in line with what previous research has found. It does makes a difference if the signal tweet is included in the data set or not. We can see that the models are not performing very well in classifying the correct tweets as positive as seen in the very low positive recall and F1 score.

4.2 Feature importance

Figure 3 shows the feature importance of the 10 most important features according to the RF model for LIWC features. The mean importance over all the features is 0.01 and the median of the set is 0.007, so actually it could be said that all 10 of these features have a higher importance than the mean importance. But it is clear that two features stand out. The 'Health' and 'Sad' LIWC categories are the most contributing to the model. And we can also see that 'Insight', 'Biological processes' and 'Anxiety' are fairly important as well. Insight includes words such as think and know. It is an collection of words used to express yourself or give insight in what you're thinking.

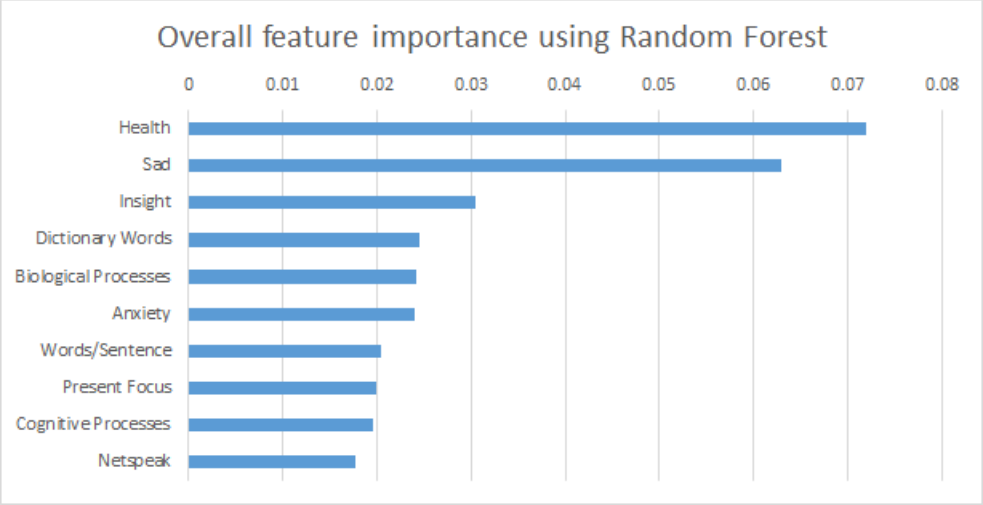

Figure 3: Feature Importance for LIWC using Random Forest

Some features that could be expected to be important such as sad and anxiety are shown to be a few of the most important features. The category 'Sad' includes words like crying and grief, and 'Anxiety' includes words such as worried and fearful.

The figures for the Linear Support Vector Machine models show 20 blue bars and 20 red bars, the blue bars are for the positive label (depressed) and the red bars are representing the negative label (not depressed).

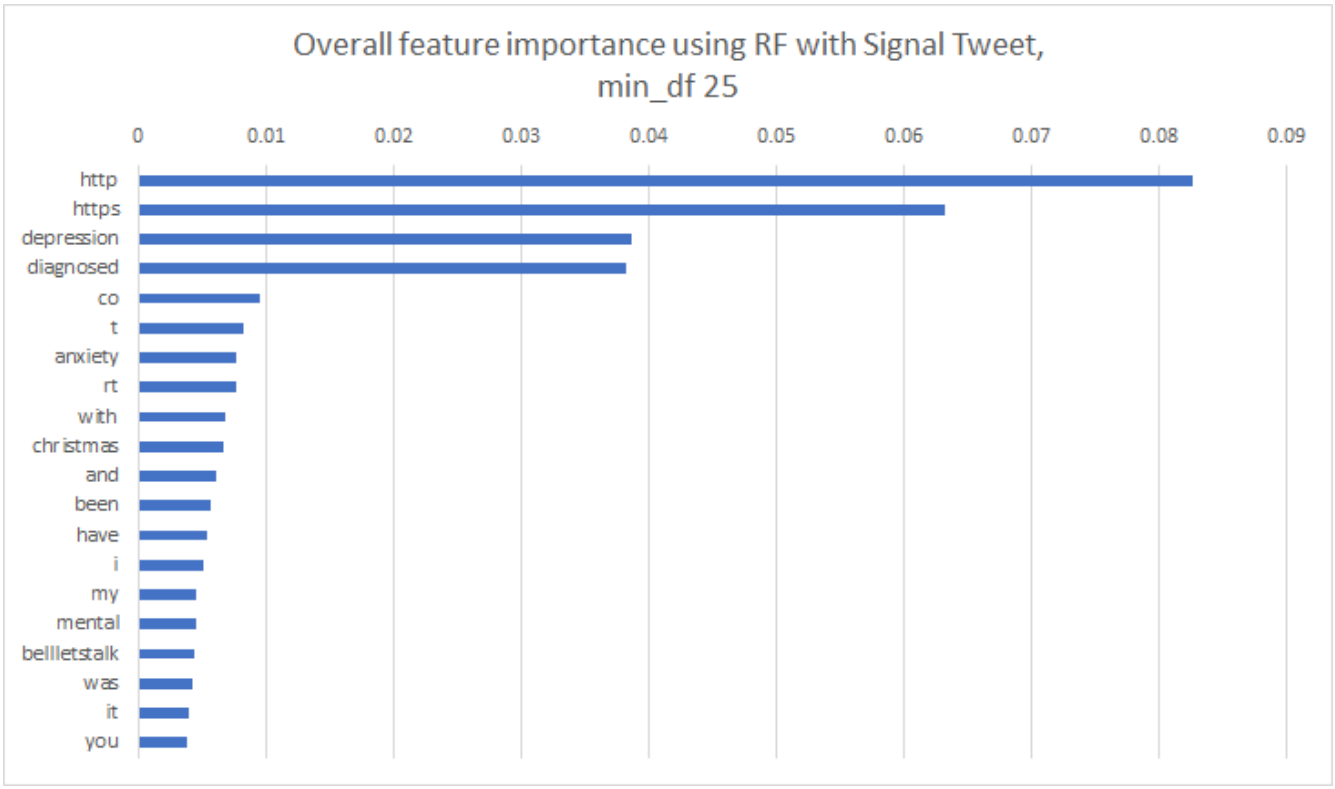

Figure 4: Feature Importance Random Forest for BOW with Signal Tweet and a min_df of 25

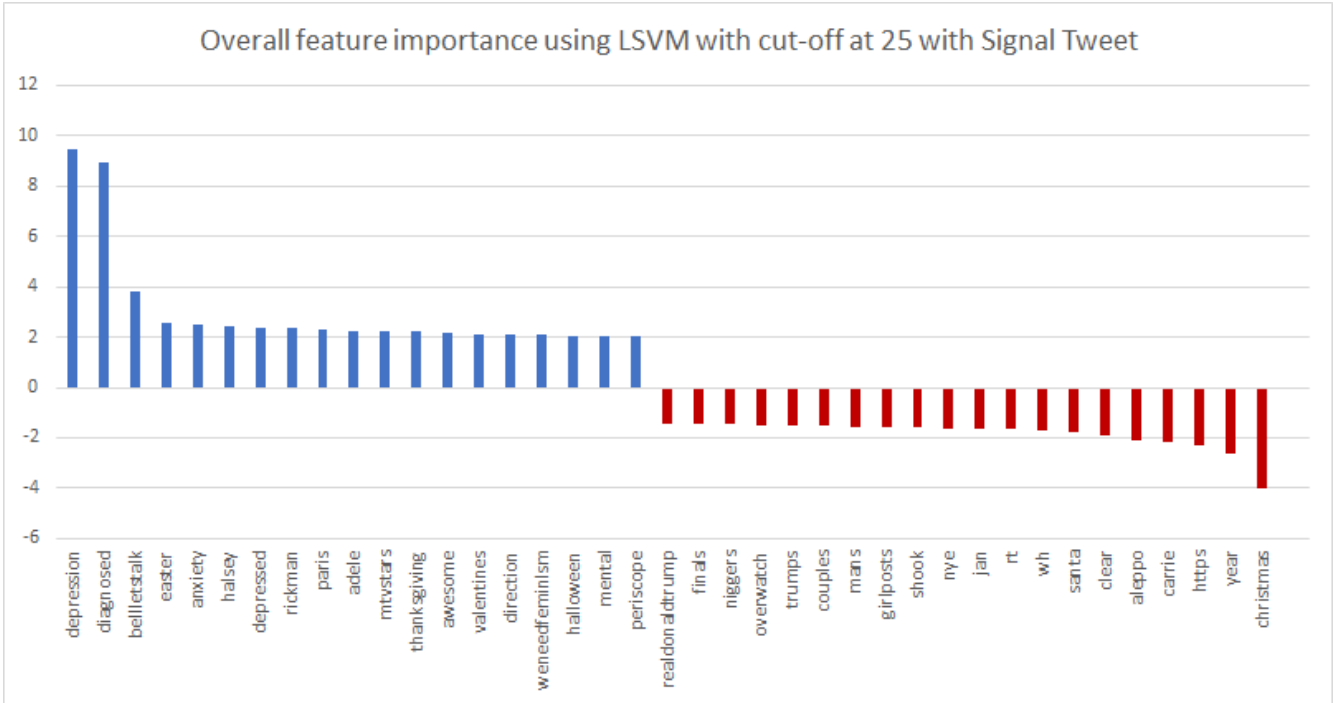

Figure 5: Feature Importance Linear Support Vector Machine for BOW with Signal Tweet and a min_df of 25

The reason the results from the BOW were so much higher than expected with both classifiers becomes clear when we take a look at the feature importances shown in Figures 4 and 5. It is apparent that with Random Forest the words diagnosed and depression are very high on the list. This is probably not a coincidence because the signal tweet on which the user was selected as depressed or not contains exactly those words.

Besides the obvious words as depression and diagnosed, some other words that relate a lot to the topic of depression signs such as 'anxiety' and 'mental' scored among the top 20 most important features for the label 'True'. 'bellletstalk' also scores very high for both models. It appears to be the hashtag used for a awareness campaign created by the Canadian telecommunications company Bell Canada to raise awareness around mental health issues in Canada.

As described previously, the same classifiers were also used to fit a model to the data set that did not contained the signal tweets. To see what kind of effect this had on the feature importance for both the Random Forest and Linear Support Vector Machine classifiers, the plots in Figures 6 and 7 were created.

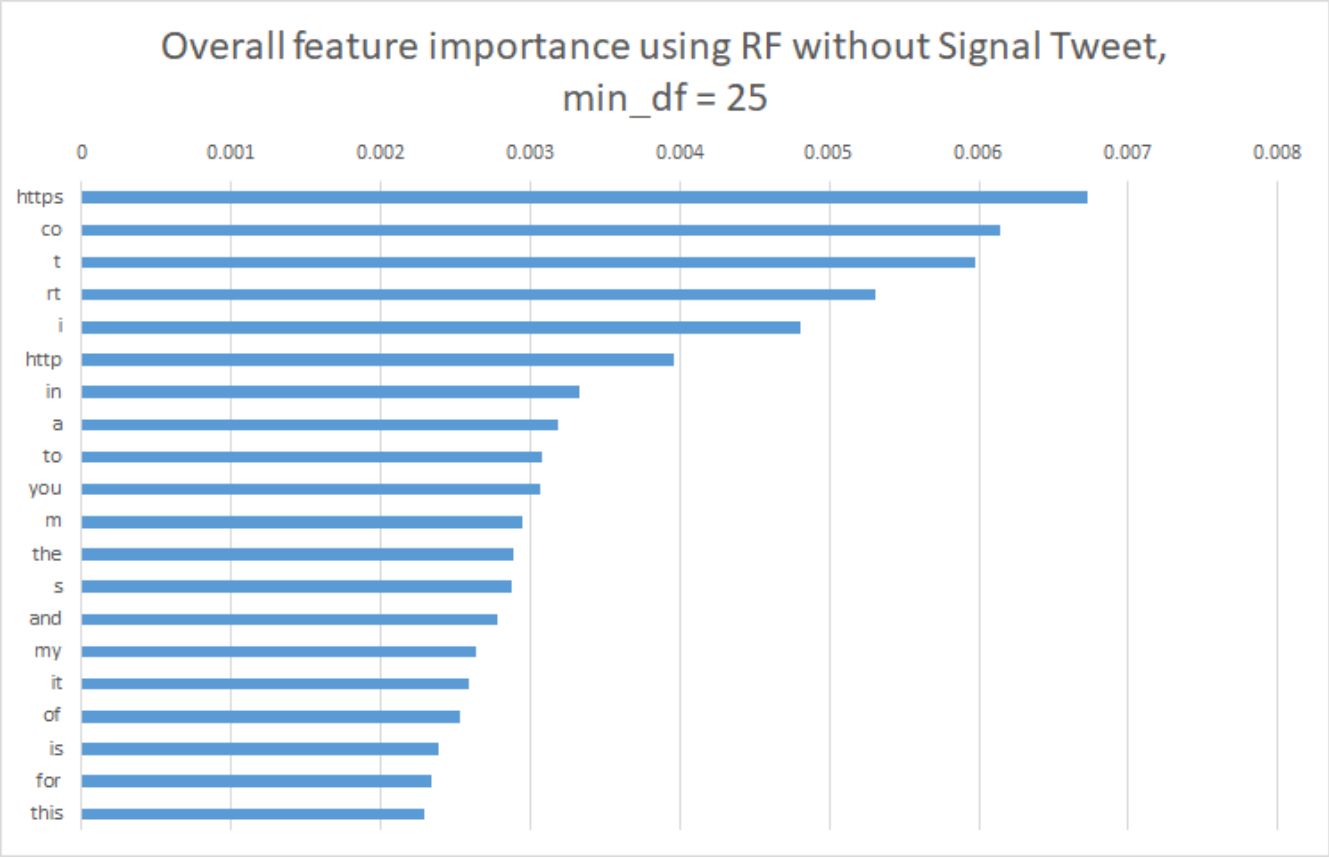

Figure 6: Feature Importance Random Forest for BOW without Signal Tweet and a min_df of 25

Looking at the scale on the x-axis in Figure 6 it is clear that the RF model without using the signal tweet have a much lower overall feature importance score compared to the RF model with the signal tweet. For example the Random Forest model with signal tweet has highest feature

importance score for http of 0.08 when using min_df 25 4 and a score of 0.09 when using 50 8. But for the Random Forest model without using the signal tweet the highest feature importance score is for https with a score of almost 0.007 in both cases 610.

It is still very noticeable how http and https and short words as rt (used for retweeting a tweet) and co are very high on the list. Http and https are used for links which can be links to websites but is mostly used for images of gifs. Co is also part of a short URL, because short urls on Twitter are in the form of "https://t.co/" followed by a string of small and capital letters and numbers. There has not been any indication by previous studies that users who might show sign of depression use a lot of URLS in their tweets, so it could be the case that the use of URLS is not a characteristic feature of users who might be depressed, but a result of different property of the users of their tweets.

There are also quite a few single letters present in the graph considering we did not use character n-grams. Looking at the tweets in which these single letters were present we found that there were different possible explanations for this. One of those was that when someone retweets a tweet that is to long to quote in its entirety, the tweet is cut of and replaced by dots. Between the last letter of the quoted tweet and the dots that replace the remainder of the tweet a space is placed like this for example: "You may be secretly longing for the hectic holiday season to s ...". Another reason for this could be that it sometimes happens that people forget the apostrophe between a word and a following 's' to indicate a possessive pronoun. And a final reason for this could be that people sometimes spell words out with spaces in between, though this happens rarely.

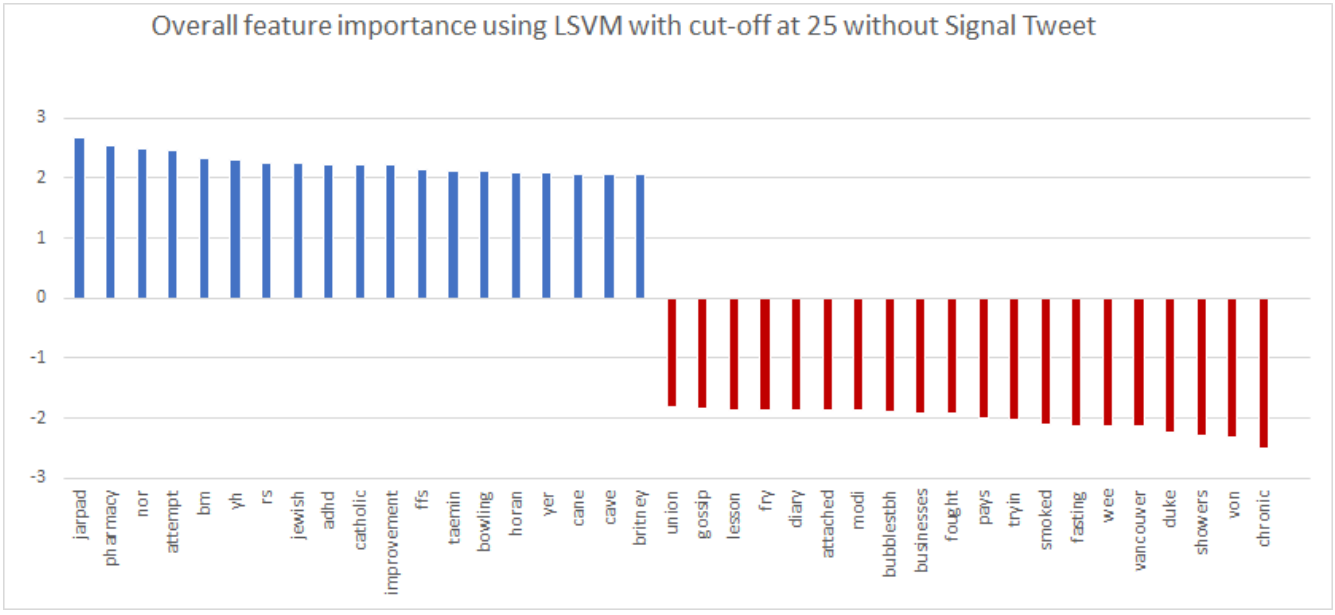

Figure 7: Feature Importance Linear Support Vector Machine for BOW without Signal Tweet and a min_df of 25

The feature importances for the Linear SVM without signal tweets seem very random at first. The difference from both the Linear SVM model with signal tweet as the RF model without signal tweets is quite large. In the Linear SVM model without signal tweets seen in Figure 7 the highest scoring feature for prediction the label 'True' are jarped, which is the Twitter handle for Actor

Jared Padeleki. For the other 19 features that are considered most important for predicting the label 'True' there is not a lot of difference in the score they got, ranging from 2.7 to 2. Comparing this to the range in score from the model with the signal tweets in Figure 5 this is quite a big of a difference. The range from the Linear SVM model with the signal tweet went from 9.5 to 2 for the top 20 features considered most important to predict the True category.

Words such as pharmacy and improvement or attempt do fit the depression context, but there is no way of checking if these words are in the context of dealing or struggling with depression symptoms from just the words alone. The words can of course be used in a different context.

The use of 'Netspeak' words is quite apparent in the features in Figure 7. Words such as 'bm', 'ffs', 'yh', 'yer' and 'rs' are abbreviations of the words 'bad manners', 'for fucks sake', 'yeah', 'your' and 'runescape'(a multiplayer online role-playing game) in most of the cases. There has not been a lot of studies that looking into the use of Netspeak in different age, gender or personality categories, but a study by HU et al [8] does state that Twitter case overall be seen as a medium where more formal linguistic styles are used. The kind of abbreviations as shows above do not fit these formal styles. So it is quite interesting that the Netspeak words are such prominent features.

4.3 Wrongly classified users

As a final evaluation measure we will take a look at the wrongly classified user/sub samples that each classifier produced. In Table 9 is shown how many predictions each classifier predicted wrong.

Table 9: The amount of wrongly classified sub samples per classifier

	LIWC whit RF	BOW with RF with signal tweet, min_df = 25	BOW with Linear SVM with signal tweet, min_df = 25	BOW with RF without signal tweet, min_df = 25	BOW with Linear SVM without signal tweet, min_df = 25
Amount of wrongly classified sub samples	2141	922	682	3451	3422
Predicted False should be True	1166	872	520	3407	2299
Predicted True should be False	974	50	162	44	1123

As expected from the results discussed in the paragraph about model performance, the models in which the signal tweets was included got the least amount of tweets wrong. It is remarkable that even though the classifier for the LIWC features included the signal tweets, the amount of incorrect predictions is more comparable to the classifier for the BOW features that didn't include the signal tweets.

To see if anything can be learned from the the wrongly classified users, the subsets with the highest confidence score for the sets that were classified as False but should be True from each classifiers will be investigated. Only the subsets classified as False while they should be True are analyzed because this is of the most interest to us. We want to know if any patterns can be discovered why certain subsets are classified as False while their confidence score indicated they should be classified as True.

Starting with the LIWC feature classifier the top 5 of users classified as False while they were supposed to be classified as True, had a confidence score of 1 for true. So even though according to the score they should be classified as True, they still got classified as False. Manually extracting those top 5 users and reviewing them carefully. Only one subset of tweets contained the signal tweet. Most of the subsets had a positive tone while one of the five stood out to be extremely negative about life. This overall positive tone could be the reason that the users were misclassified.

Looking at the top 5 wrongly classified subset for Random Forest with the signal tweet all sets had a confidence of 1 for the label true but did got classified as False. Linear SVM had a confidence score ranging from 6 to 2. The first thing that stood out was that both RF and Linear SVM had 1 subset in their top 5 that contained a signal tweet. The subset in RF had an overall positive tone while the subset in Linear SVM contained only tweets directed at a famous youtuber with words related to depression, self harm and almost a cry for help or attention. A few more subsets in both classifiers contained words related to self harm and negativity in general. A lot of the tweets in the subsets contained quite a few links to different sites such as Tumblr for images or retweet links from Twitter itself.

for Linear SVM we noticed a much lower confidence score without the signal tweets than with the signal tweets. The top 5 had a confidence score ranging from 1.5 to 1.1. The top 5 highest confidence score for the wrongly classified tweets for Random Forest had two subsets that had a confidence score of 1 and three that had a confidence score of 0.99. The content of the text was predominantly positive tweets with a mix of complaining about life and discussing TV series. There were really not a lot of words that relate to depression or anxiety. The only word that came close in these subsets was stressed. This misclassification could suggest that we are not actually classifying if a user might show sign of depression, but that instead a different personality type, age or gender might be used to classify on.

It is quite difficult to give a clear reason why these users were misclassified. One of the reasons could be that some people use words that, according to the models, would indicate that they could be showing signs of depression in just one of their tweets while having a very positive tone in the majority of their tweets. This could cause the users to be misclassified. But this was predominantly seen in the models that did include the signal tweets. The models without the signal tweets showed barely any connection to depression in the use of words that we would expect, nor were there any apparent patterns to be found. This could suggest that te model might be training on different pattern than we expected.

5 Discussion

The results have shown that looking at pure accuracy scores doesn't always say everything. If we were to only consider accuracy and the classification report we might conclude that we should include the signal tweet an choose the Linear Support Vector Machine in combination with our Bag of Words feature set. Because if we compare this model to our optimized Random Forest classifier for the LIWC features the numbers might suggest that the Linear SVM BOW classifier has a much higher accuracy, Positive Recall and F1 score.

But these surprisingly high scores for the Linear SVM suggested that something made the classification almost too easy. Looking at the feature importances we can understand why. The most important features for both BOW models included the word 'depression' and 'diagnosed'. If we take into account that the signal pattern on which the signal tweets were chosen did include those words, we might conclude that because the signal tweets were still included in the data set the problem has been made to easy and therefore might not have learned anything.

We can see from the results that removing the signal tweets did have a big impact on the accuracy scores for the models with BOW features. The highest accuracy score dropped from 94% to 71% which is more in line with what most other studies show. But if we take a look at the Positive recall and F1 score we can see that these models do not perform well in predicting the True label correctly. Where Linear SVM shows an already very low score of 0.418 and 0.491 respectively, the Random Forest model scores even lower. It is clear that Random Forest is not a suitable classifier for dealing with BOW features. The Random Forest classifier classified almost everything as false and achieved a very low score for the positive recall of 0.138 and 0.240 for the F1 score.

If we would to compare LIWC and BOW features based on the model accuracy for the data set that included the signal tweet, we can conclude that the BOW with a Linear SVM classifier achieved the highest result but might not have learned a lot of useful information. Because we did not try to analyze our data set without the signal tweet by LIWC we cannot make any statements comparing the LIWC features with the BOW features in this case. But we can compare the BOW feature with and without the signal tweet and from this we can come to the point that without the signal tweets, the BOW classifier might not be the best model for classifying subsets of tweets as possibly showing signs of depression when there are no words such as depression directly involved.

The feature importance gave us a few very interesting results. It showed us first of all, as discussed above, that including the signal tweet had a very big effect on the BOW feature models. The words 'depression' and 'diagnosed' has some of the highest importance when using Random Forest as well as the Linear Support Vector Machine. The highest scoring features without using the signal tweets included celebrities like Jared Padalecki ('whose Twitter handle is 'jarped') and Niall Horan for the Linear SVM. This might suggest a certain audience category on which the Linear SVM classifies the users, namely people who are interested in the tweets of Padalecki and Horan, which could be the fans of the content they produce. The Random Forest really focuses in urls as shown by the high importance of 'https' 'co' and interactions as shown by 'rt'.

The feature importance using the LIWC features seem more in line with what research shows about

online expression of depression. The high importance of health and sad can be explained by words as depression, diagnosed which come from the signal tweet. But categories such as Insight and Present focus seem so confirm to what the research of Park et al [14] stated. People who suffer from depression are more likely to express their feelings and their thoughts.

The research paper of Shen et al [18] who constructed the data set that we used, caused a lot of questions. The statement made in their paper about the 'strict' pattern, implied that they were talking about a contiguous pattern. The pattern was not that strict or contiguous and allowed for different words in between the words "diagnosed" and "depression" which they did not make clear. This appeared not to cause any inconsistencies in the data and might be the right approach to select more signal tweets, but could be better explained.
There were also some contradictions between the data they described and the data they delivered. The data set they provided did not seem to contain the same information as they specified in their paper, in terms of quantity. This might be because of some errors while downloading, but the downloading of this data set was tried several times and every time with the same data as a result.

6 Conclusions and Further Research

The goal of this research was to detect if a person might have depression signs based on Twitter text data using classification models. To do so, two different sets of features were compared: features generated by the LIWC tool and a Bag of Words feature set. Different classifiers were compared as well to find the best performing model.

Comparing the best performing LIWC models and BOW models based only on accuracy score and the classification report showed that the BOW models had a much higher accuracy, positive recall and positive F1 score. This raised some questions, because an accuracy of 94% was a lot higher than found in previous research that tried the same thing using BOW features. So see what might have caused this, the most important features according to the feature importance score were explored. This showed that in the BOW model the words that got the highest score were 'depression' and 'detection' which are also the most important words in the pattern on which signal tweets were selected. So to see if this indeed caused the high accuracy score we removed the signal tweets from the data set and found that the same model for BOW features got an accuracy score of 71%. These results were comparable with previous research. However, looking at the positive recall and F1 score the model might not be that accurate at predicting if someone might be showing signs of depression based on a subset of their tweets.

Because LIWC categorized the words, it might tell is more about the kind of words are most important to classify a user as 'depressed', than than the BOW model which shows importance for specific words. We should keep in mind that for the LIWC models we did include the signal tweets, so we cannot compare it directly to the BOW model without the signal tweet included.

Looking back at the results from the BOW features we can ask ourselves if we are actually classifying users as depressed or if we are classifying users on a specific age, gender and/or personality. The abbreviations that are considered important features according to the BOW feature model are mostly used by a audience that uses Twitter more informal. The same goes for the names of celebrities that have a high feature importance. The use of the names of these celebrities are connected to an audience that might watch shows where they appear of listen to their music, in other words are fan of their content. This might be a result of sampling bias. For feature work we might look into the process of collecting samples on Twitter for predicting depression.

Another possible feature research to investigate could be the effect of sub samples of tweets taken per user. In this study we decided to take 20 tweets per user as a sub samples, and five sub samples per user in total. By changing these numbers we might see different results in the accuracy and important features. And finally, a combination between using BOW and LIWC features could be examined. A lot of research is already looking into a combination of different features, so combining LIWC an BOW could be an interesting addition.

References

[1] Domo resource - data never sleeps 7.0. https://www.domo.com/learn/data-never-sleeps-7, 2019. [Online; accessed 17-July-2020].

[2] Bazarova, N. N., Choi, Y. H., Sosik, V. S., Cosley, D., and Whitlock, J. Social sharing of emotions on facebook. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW 15* (2015).

[3] Berryman, C., Ferguson, C. J., and Negy, C. Social media use and mental health among young adults. *Psychiatric Quarterly 89*, 2 (2017), 307–314.

[4] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16* (Jun 2002), 321–357.

[5] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. Clpsych 2015 shared task: Depression and ptsd on twitter. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (2015).

[6] Dirkson, A., Verberne, S., Sarker, A., and Kraaij, W. Data-driven lexical normalization for medical socialmedia. *Multimodal Technologies and Interaction 3*, 3 (2019).

[7] Gamon, M., Choudhury, M., Counts, S., and Horvitz, E. Predicting depression via social media. *Association for the Advancement of Artificial Intelligence* (2013).

[8] Hu, Y., Talamadupula, K., and Kambhampati, S. Dude, srsly?: The surprisingly formal nature of twitter's language. In *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013* (2013), AAAI press, pp. 244–253. 7th International AAAI Conference on Weblogs and Social Media, ICWSM 2013 ; Conference date: 08-07-2013 Through 11-07-2013.

[9] Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., and Ulhaq, A. Depression detection from social network data using machine learning techniques. *Health Information Science and Systems 6*, 8 (2018).

[10] Kelly, Y., Zilanawala, A., Booker, C., and Sacker, A. Social media use and adolescent mental health: Findings from the uk millennium cohort study. *EClinicalMedicine 6* (2018), 59–68.

[11] Losada, D. E., Crestani, F., and Parapar, J. erisk 2020: Self-harm and depression challenges. *Lecture Notes in Computer Science Advances in Information Retrieval* (Apr 2020), 557–563.

[12] Nguyen, T., O'Dea, B., Larsen, M., Phung, D., Venkatesh, S., and Christensen, H. Using linguistic and topic analysis to classify sub-groups of online depression communities. *Multimedia Tools and Applications 76*, 8 (2015), 10653–10676.

[13] Ophir, Y., Asterhan, C. S., and Schwarz, B. B. Unfolding the notes from the walls: Adolescents' depression manifestations on facebook. *Computers in Human Behavior 72* (2017), 96–107.

[14] Park, M., McDonald, D., and Cha, M. Perception differences between the depressed and non-depressed users in twitter. *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013* (01 2013), 476–485.

[15] 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., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[16] Pennebaker, J., Francis, M., and Booth, R. Linguistic inquiry and word count (liwc).

[17] Rinke, M., Bundy, D., and Stein, R. Increasing recognition and diagnosis of adolescent depression: Project redde: A cluster randomized trial. *Pediatric Quality and Safty 4*, 6 (2019).

[18] Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., and Zhu, W. Depression detection via harvesting social media: A multimodal dictionary learning solution. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (2017).

[19] Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., and Zhu, W. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (2017), pp. 3838–3844.

[20] Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H. Recognizing depression from twitter activity. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI 15* (2015).

[21] Depression fact sheet. https://www.who.int/news-room/fact-sheets/detail/depression.

Appendix

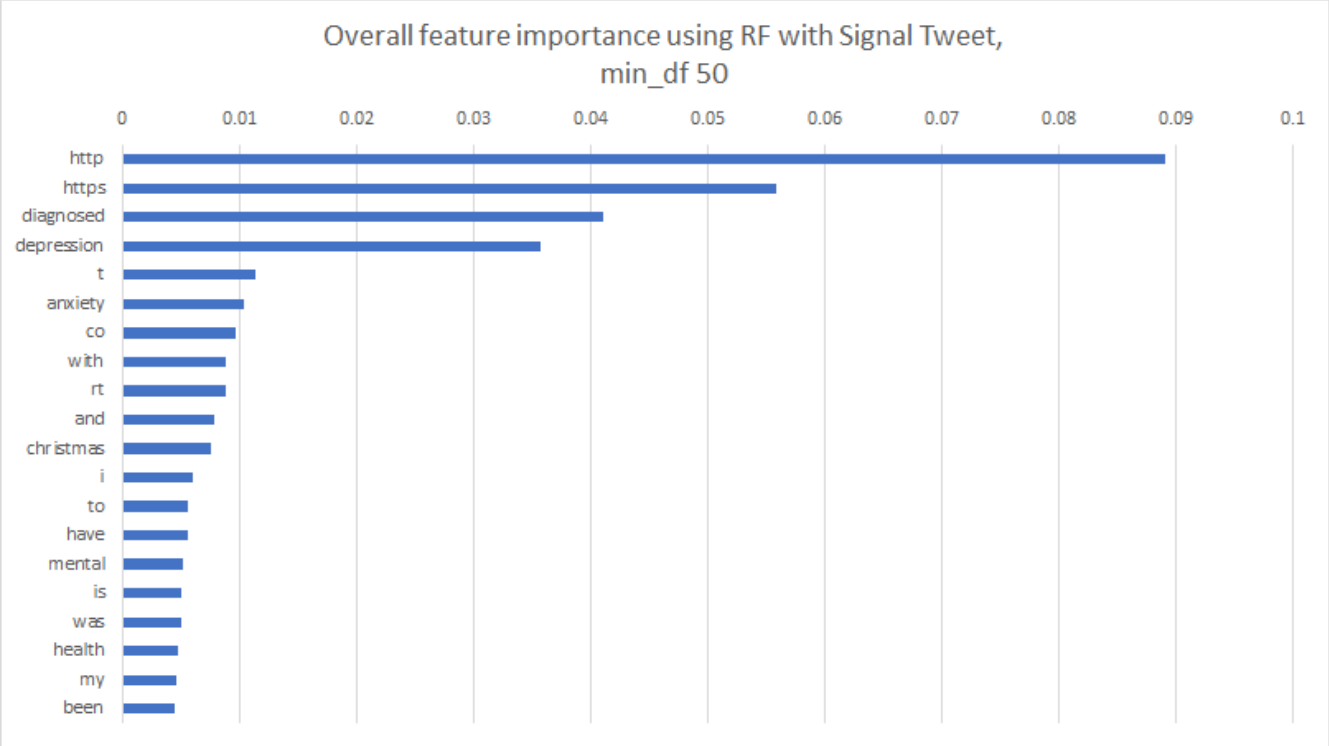

Figure 8: Feature Importance Random Forest for BOW with Signal Tweet and a min_df of 50

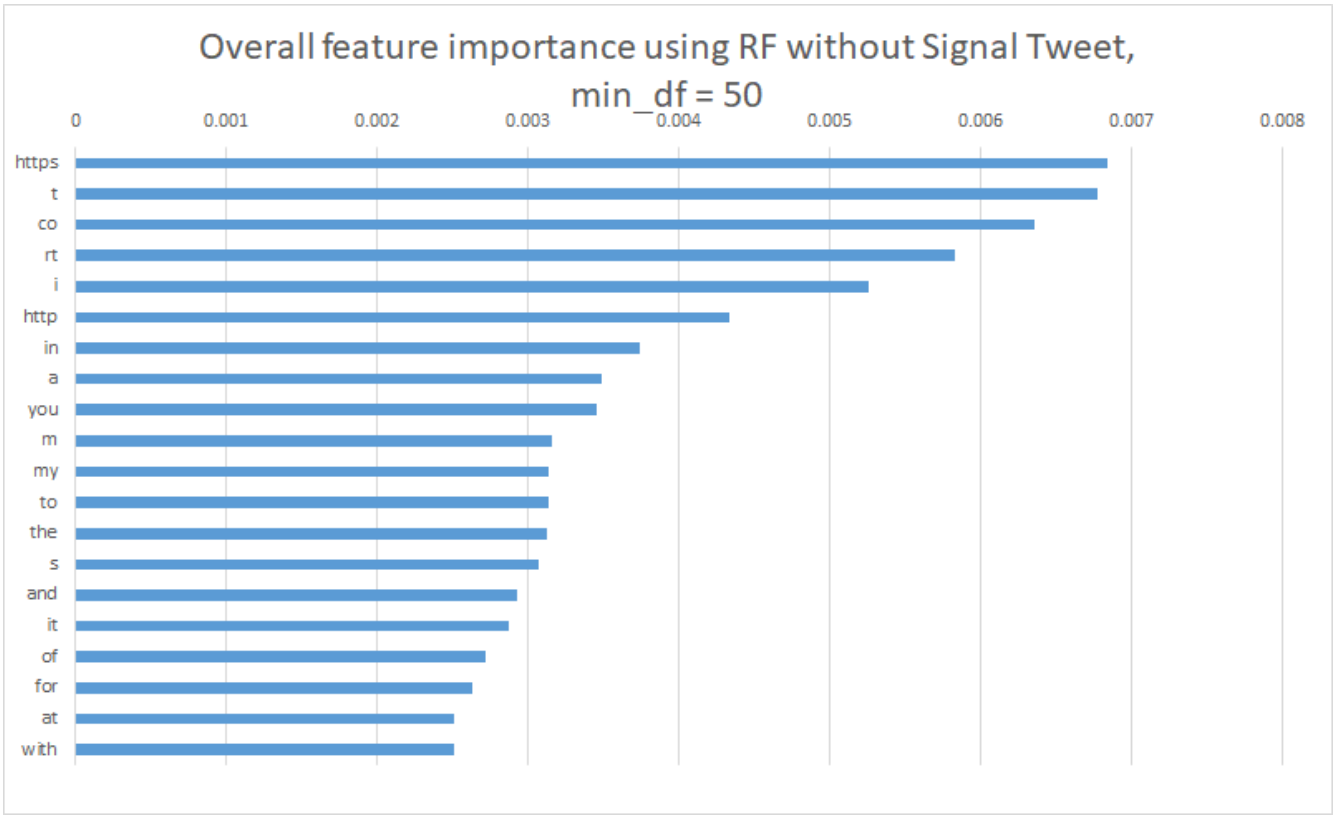

Figure 9: Feature Importance Random Forest for BOW without Signal Tweet and a min_df of 50

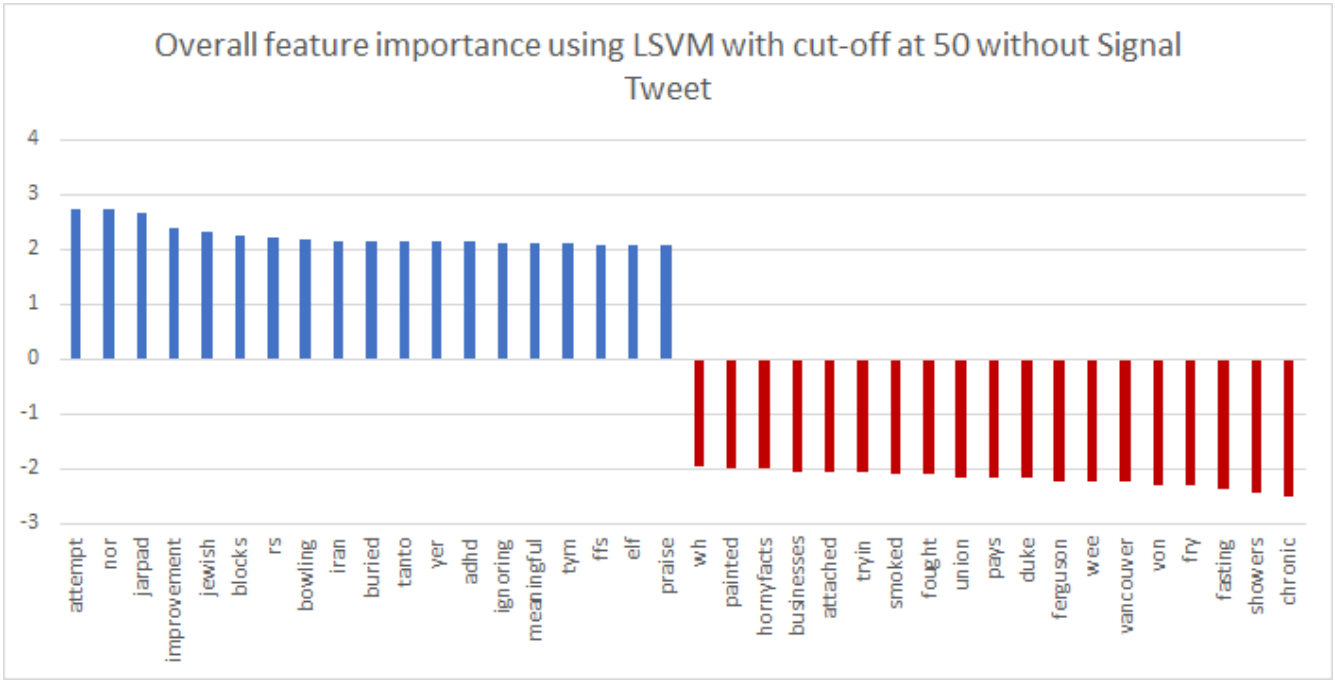

Figure 10: Feature Importance Linear Support Vector Machine for BOW without Signal Tweet and a min_df of 50

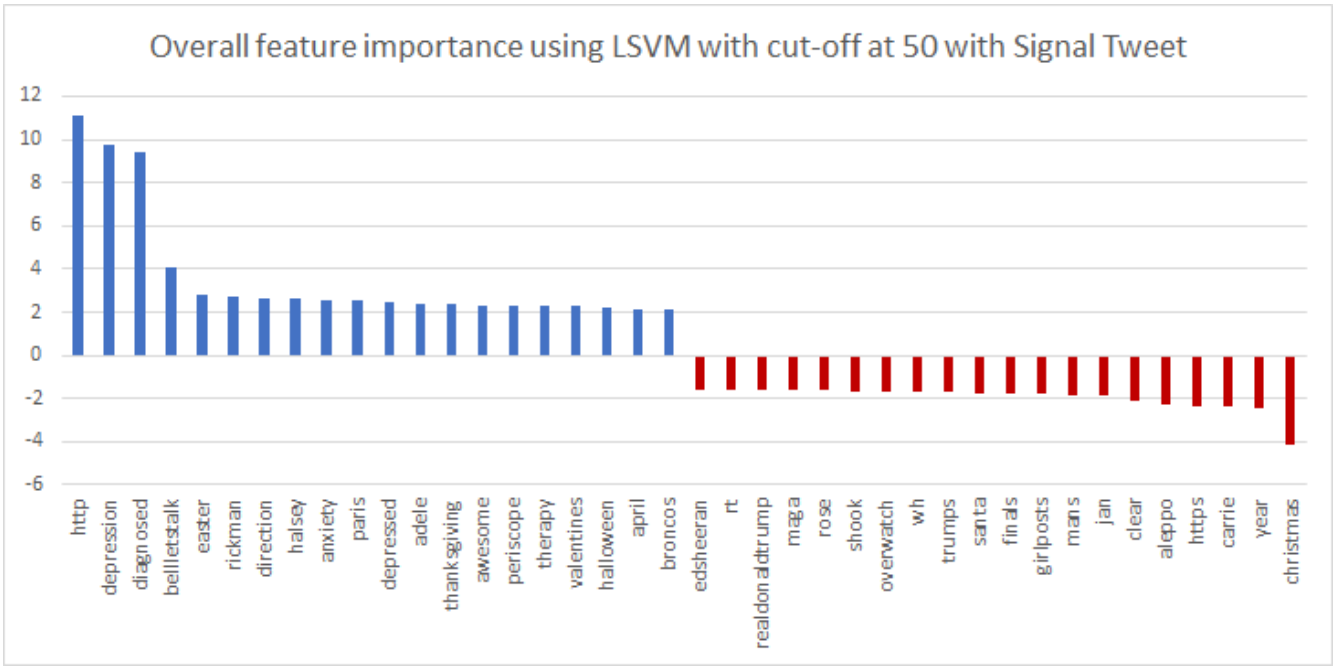

Figure 11: Feature Importance Linear Support Vector Machine for BOW with Signal Tweet and a min_df of 50

Table 10: The amount of wrongly classified sub samples per classifier with min_df=50, using BOW

	BOW with RF with signal tweet, min_df = 50	BOW with Linear SVM with signal tweet, min_df = 50	BOW with RF without signal tweet, min_df = 50	BOW with Linear SVM without signal tweet, min_df = 50
Amount of wrongly classified sub samples	880	737	3435	3640
Predicted False should be True	824	550	3380	2462
Predicted True should be False	56	187	55	1178