Bachelor Thesis 7.5 EC – Artificial Intelligence

Predicting balanced participation in two-person online conversations

Nikki Moolhuijsen
n.s.moolhuijsen@students.uu.nl
Student number: 6583296

Supervisor: Anna Wegmann
a.m.wegmann@uu.nl

Second reader: Leendert van Maanen
l.vanmaanen@uu.nl

Faculty of Humanities
Utrecht University
The Netherlands
02/07/2021

Contents

1. Introduction 2
2. Related Work 3
3. Feature Selection 4
   - 3.1 Method 4
     - 3.1.1 Data 4
     - 3.1.2 Features 5
     - 3.1.3 Logistic Regression 6
   - 3.2 Results 7
   - 3.3 Discussion 9
4. The Models 10
   - 4.1 Method 10
   - 4.2 Results 11
   - 4.3 Discussion 11
5. Conclusion 12
6. References 13

1. Introduction

Social media refers to online interactions between people in online communities, partaking in activities such as sharing information and discussing ideas. Currently, 4.14 billion people use social media, and that number increases with an average of 2 million every day (Kemp, 2020). The average time spent on social media is 2 hours and 32 minutes per day per person (Peterson, 2021).

Research has shown that the use of social media can have huge impacts on life, both negative and positive (Berry et al., 2018; Berryman et al., 2018). Toxic behaviour, such as cyberbullying, hate speech and abusive language, can be present on online platforms, and can affect mental health in a negative way (Alhajji et al., 2019; Bhat, 2016). Differently, online communities that offer support and information on mental illness may reduce feelings of loneliness and improve mental health. (Glazzard & Stones, 2021; Wadden et al., 2020)

Toxic behavior, among other things, affects the quality of a conversation. Investigating what contributes to quality of a conversation may improve both the conversation itself and online communities overall. The negative mental health effects around toxic behavior has prompted research into detecting and reducing this type of behavior (Aroyehun & Gelbukh, 2018; Singh et al., 2020). A different type of behavior that contributes to quality is prosocial behavior, which can be defined by actions that benefit another individual (Eisenberg & Mussen, 1977), for example offering support. Moreover, prosocial behavior is positively associated with increased well-being (Martela & Ryan, 2016)

Likewise, research into conversation quality may also be beneficial to artificial intelligence. Chatbots have become fairly common, for example through the integration of chatbots into customer service. Social chatbots such as Woebot (Martin, 2018) and ELIZA (Martin, 2018), are designed to engage in conversation with human, where mimicking human conversations is one of the goals set out for the developers. Specifically, ELIZA was developed to imitate a therapist, whom people could ask questions to. Woebot, like ELIZA, also falls into the category of a mental health chatbot, and is used by people to improve their mental health. The investigation of the quality of online interactions may give further insight into making chatbots more human-like.

Quality is a very broad term and thus can be defined in a number of ways. What contributes to the quality depends on the type of conversation. In discussion based communities, where people go to discuss a specific subject, providing correct information or giving solid arguments might be important to keep the discussion going. In support communities, people share their experiences and feelings and in return ask for advice or they just want to be heard. The quality of these types of conversations may be defined by how friendly or supportive the comments are. Other, more light hearted communities, may be used for entertainment purposes. In these communities, laughter and telling jokes, among other things, may be indicative of a good quality comment or conversation.

Something else that might contribute to the quality is how much each person adds to the conversation. In a two-person conversation, discussing a topic can only happen when both parties are actively involved in the conversation. Explaining your view and giving arguments, providing counter arguments or showing agreement or disagreement are essential for sustained discussion.

In other types of conversations, active involvement may be sharing experiences or feelings or telling a story. The quality of a conversation may improve when all participants contribute equally, or reduce if some participants have nothing to say. The term that will be used to describe equal contribution by all participants to a conversation is balanced participation. But what exactly entails balanced participation? There are different ways to define balanced participation, such as asking the same amount of questions, or giving equal amounts of feedback. This thesis will use equal number of words as the definition for balanced participation. Actively adding to a conversation is often linked to the amount of words spoken, for sharing ideas and opinions is hardly possible when using few words. Moreover, a study on conversation killers (i.e., features that are associated with posts that are unlikely to be replied to), suggests that increasing the number of words in a post will increase the probability of that post being replied to (Jiao et al., 2018).

The goal of this thesis is to understand what 'good' quality conversations have in common. Specifically, what do conversations with balanced participation in terms of equal share of words have in common? The research question that this paper will try to answer is "Which features are significantly associated with balanced participation in online communities?" This is done by i) preparing a data set of two person conversations (Section 3.1.1), ii) develop methods to identify features in conversations, see Section 3.1.2 for a list of all features, iii) use logistic regression on these features (Section 3.1.3), iv) make three models with different features (Section 4.1), and v) use logistic regression on all three models to predict balanced participation.

The features that were shown to be significantly associated with balanced participation are: bigrams, overlap of words, ratio of FPSPs, ratio of SPPs, length, first person singular pronouns, second person pronouns, ratio of questions and conjunctions. The second model of the three performed the best, and consisted of 8 of the 9 features mentioned above. Model 2 was an extension of Model 1, and showed that adding more features increased the performance. The model could be improved by adding more and better features. Balanced participation is a contributing factor to the quality of a conversation, and so finding features that are associated with balanced participation gives further insight into what affects the quality of a conversation.

2. Related Work

Behavior detection. Toxic behavior comes in many different shapes and forms, and can cause serious mental health issues. Research has been done to identify these toxic behaviors, such as cyberbullying (Huang et al., 2018) and aggression (Aroyehun & Gelbukh, 2018)

Other types of behavior have also been a topic of research. The impact of politeness was determined by identifying linguistic features of politeness (Burke & Kraut, 2008; Jurafsky et al., 2014). A study on gratitude in online communities focused on building a model for gratitude, which the platform can use to encourage gratitude throughout the community (Makri & Turner, 2020). Furthermore, prosocial outcomes were predicted from the initial comment of a conversation, using conversational features (Bao et al., 2021).

Additionally, other types of behavior such as sarcasm (Davidov et al., 2010) and attitude (Radev et al., 2010) was studied to identify these in online conversations.

Conversation quality. In an attempt to improve the quality of comment threads under news articles, "good" conversations, called ERICs, were identified in these comment threads. An ERIC stands for Engaging, Respectful, and/or Informative Conversation. They hypothesized that identifying and encouraging these types of conversations will make the community more civil and encourage participation (Napoles, Pappu, et al., 2017; Napoles, Tetreault, et al., 2017). Moreover, another study focused on developing a method to classify the quality of blog comments (FitzGerald et al., 2011).
A study on the quality of a conversation in chatbots focused on balance. Four aspects of conversation were controlled to understand the effect they have on the quality of a conversation (See et al., 2019).

3. Feature Selection

3.1 Method

3.1.1 Data

The conversations used in this paper came from the social media platform Reddit (Huffman & Ohanian, 2005), a website containing many different communities where people can post about and discuss the specific topic of that community. Such a community is called a subreddit.

Cornell Conversational Analysis Toolkit (Chang et al., 2020) is a toolkit for analyzing conversations. Included in this toolkit are different corpora, a corpus being a collection of conversations. A conversation is made up out of a post with a comment thread (i.e., a sequence of comments replying to one another). These comments or posts are stored in objects called 'utterances'. The corpus that was used is 'reddit-small-corpus', a collection of conversations from 100 highly active subreddits, with 100 comment threads taken from each subreddit. In total this corpus has 8286 posts and 288846 comments under these posts.

This thesis will look at two-person conversations, however, the comment threads from the 'reddit-small-corpus' corpus consist of comments made by multiple (more than two) speakers. Every utterance in the corpus contains the username of the speaker, which was used to extract the two-person conversations. A sequence has to have a minimum length of four to be considered a conversation. This number was chosen since it would take at least 2 comments per speaker for there to be an actual back and forth in the conversation. Any number below four would result in at least one of the speakers only commenting once, which does not suggest balanced participation.
Reddit has a function that lets you directly reply to part of an earlier comment by quoting that part in your comment. Since these quotes are not spoken by the speaker of the comment they're in, but rather an addition to the conversation to make clear what exactly is being responded to, they should be removed from the comment. An utterance also contains the text of the post or comment, but sometimes these texts were empty, meaning even though there was a reply, nothing was actually written. Any conversations that had less than four comments after removing the utterances without text were not used. From the 17090 two-person conversations 453 conversations were removed due to utterances without text.

3.1.2 Features

Category	Example words	Category	Example words
Positive emotion	love, sweet, nice	Discrepancy	should, would, could
Negative emotion	hurt, ugly, nasty	First-person singular	I, me, mine
Swear words	damn, piss, fuck	Second person	you, your, thou
Tentative	maybe, perhaps, guess	Perceptual processes	observing, heard, feeling
Insight	think, know, consider	Adverbs	very, really, quickly
Conjunctions	and, but, whereas	Interrogations	how, what, when
Causation	because, effect, hence	Assent	agree, okay, yes
Negations	no, not, never		

Table 1: Table of LIWC categories used as features with some example words.

LIWC (Tausczik & Pennebaker, 2010), a text-analysis program, has a dictionary of words that have been placed into one or more linguistic or psychological categories. The categories that were used as features can be found in Table 1. The words that fall into a certain category were counted for every conversation.
These features were chosen to investigate due to the possibility that they may encourage participation in the conversation.

The categories positive emotion, negative emotion and and swear words may set the tone of the conversation. A positive conversation may be enjoyable and thus encourage participation, whereas a conversations that seems more negative might discourage someone to participate further since it might not be as enjoyable. Furthermore, swear words can have a negative effect on participation, because one might take offense to these swear words.

The following categories contain words used in types of comments that could imply interest in the conversation and therefore active participation. Such types of comments could be, for instance, telling a story or discussing a subject. The categories adverbs and perceptual processes contain words one may use when sharing feelings or describing experiences. Tentative and insight contain words used when giving opinions or information. Discrepancy has words that can be used for giving advice. Assent consists of words that show agreement, whereas negations may show disagreement. Both of these categories may guide the path the conversation will take. Agreement might make the conversation more pleasant which will stimulate participation, on the other hand, disagreement might encourage further discussion. The categories conjunctions and causation are composed of words one may use in discussions when giving arguments or explaining things. Interrogations has words that can be used when asking questions.

Pronouns may also be interesting to look into. First person singular pronouns point to participants speaking about themselves, which could strengthen the connection between the speakers. Second person pronouns may indicate an interest in the other speaker, such as asking questions.

Besides LIWC categories, other features that were looked at are:

Length of the conversation. The total number of comments in the conversation. The longer the conversation is, the more interested both speakers may be in the conversation. Otherwise the conversation would probably have been cut short.
Laughter. The total number of instances of laughter. A regular expression (RE) for laughter was used to identify instances of laughter, which is the formula: [aA]*[hH]+a+h+a+(h+a+)*?h*. This

RE identifies words such as 'haha' and 'haaahaha' amongst other variations. The regular expression comes from (Bao et al., 2021).
Laughter may indicate interest in the conversation. If the conversation is fun, it will be more likely the conversation will continue.
Questions. The total number of question marks in the whole conversation. Asking questions could allude to interest in the other speaker or wanting information.
Ratio of questions. The ratio of question marks in the comments of the first speaker to question marks in the comments of the second speaker. Balance in the amount of questions being asked may imply balance in the conversation overall.
Overlap of words. Cosine similarity between words in all comments of the first speaker and the words in all comments of the second speaker. Cosine similarity is a metric to compare two sets of objects, in this case a set of words. Not all words in the conversation are used, but rather only words specific to that conversation. Words such as pronouns, articles and prepositions are not compared. This metric counts the instances of topic-specific words and transforms them into a vector. For both speakers a vector is made, and the cosine of the angle between these vectors is the value for this feature.
Bigrams. Cosine similarity between bigrams of the all the comments of the first speaker and bigrams of all the comments of the second speaker.
Ratio of first person singular pronouns (FPSPs). The ratio of FPSPs in all comments of the first speaker to FPSPs in all comments of the second speaker.
Ratio of second person pronouns (SPPs). The ratio of SPPs in all comments of the first speaker to SPPs in all comments of the second speaker.

3.1.3 Logistic Regression

What is logistic regression? Logistic regression is a statistical model which classifies input variables into a class, in binary logistic regression, these classes are labeled 0 and 1. A sigmoid function is fit to the training data, and shows the probability that an input variable is classified as 1, based on the independent variables of the input. If the probability is greater than 50 percent, the input will be classified as 1, if it is lower than 50 percent, it will be classified as 0. The function is fit to the data using 'maximum likelihood', a method that determines which parameters are most likely to produce the data.

The Feature Selection Model

Once the two-person conversations are extracted from the corpus, these conversations are labeled either balanced (1) or unbalanced (0) by calculating the ratio of words of the two speakers. A conversation is labeled 1 if the ratio is a number between 0.9 and 1.1. From the 16637 two-person conversations, there are only 124 conversations with a word ratio of 1. This is not enough data to run the model on, so a range of 0.9 to 1.1 is used instead to increase the data set. With this range, there are 1432 conversations labeled 1. Conversations are labeled 0 if the word ratio is higher than 3 or lower than 0.3. This ratio is quite extreme, and is chosen for that reason. Conversations with a word ratio of, for example 0.7, might not differ much from balanced conversations in terms of the features that are investigated. Whereas two sets of conversations with very different word ratios may differ much more in terms of the features. The bigger the difference between the two sets, the better the model will perform.

In total there are 1432 balanced conversations and 3698 unbalanced conversations. Logistic regression works best when the data has an equal amount of data points classified as 1 and data points classified as 0. Running the model on this uneven data set causes the model to be biased towards the class with more data point. Table 2 shows the results of a preliminary model that

Precision	0.625
Recall	0.190
F-score	0.291

Table 2: Precision, recall and F-score of the preliminary model.

was run on the data set with all 5130 conversations. Recall of this model is very low due to bias towards the class with label 0: of the 185 balanced conversations, only 35 were predicted as balanced. A low score on recall causes the f-score, a measure of the model's total accuracy, to drop. Therefore, it is chosen to work with an even data set: 1432 balanced conversations and 1432 unbalanced conversations. From the 3781 unbalanced conversations, the first 1432 conversations are used and added to the 1432 balanced conversations to create the final data set with a total of 2864 conversations.

The data was then split into a train, validation and test set with a ratio of 72/13/15 respectively. The model contains all the features described in Section 3.2 and is trained on the train set and tested on the validation set. The p-value and coefficient of a feature describe the relationship between the independent variable (feature) and the dependent variable (class). These values will lead the selection process for the three final models.

3.4 Results

Accuracy	0.66
Precision	0.65
Recall	0.66
F-score	0.66

Table 3: Accuracy, precision, recall and F-score of the feature selection model.

The p-values and coefficients can be found in Table 4. The p-value shows the statistical significance of the feature, where $p<0.05$ is considered significant and $p<0.001$ is statistically highly significant. The coefficient shows how much the mean of the dependent variable changes when the independent variable changes and all other independent variables stay the same. A coefficient greater than 1 shows that as the value for the feature increases, the mean of the dependent variable may also increase with probability equal to the coefficient.

For example, the feature length has a coefficient of 1.10. This means that if the value of the feature is increased by one, then the probability that that conversation is labeled 1 increases with 10 percent. If the coefficient is below 1, the probability decreases.

The p-values show that eleven features are statistically significant: positive emotion, bigrams, perceptual processes, ratio of questions, conjunctions, first-person singular pronouns, second person pronouns, overlap of words, ratio of FPSPs, ratio of SPPs. These lasts seven all have a p-value equal to or below 0.001, meaning they are highly significant. Moreover, the features overlap of words and bigrams have high coefficients: 46.7 and 22.7 respectively.

The model had an F-score of 0.66, which can be found in Table 3.

Feature	P-value	Coefficient	Feature	P-value	Coefficient
Positive emotion	0.0288	0.979923	Discrepancy	0.4428	1.010067
Negative emotion	0.7776	0.997337	First-person singular	0.0001	1.076799
Swear words	0.2887	1.026150	Second person	0.0001	1.037846
Tentative	0.8207	1.002486	Perceptual processes	0.0083	0.972877
Insight	0.1084	0.985674	Adverbs	0.3224	1.009632
Conjunctions	0.0001	0.967041	Interrogations	0.9994	1.000015
Causation	0.2963	1.010680	Assent	0.7530	1.019743
Negations	0.1617	1.018846			
Length	0.0000	1.095900	Questions	0.0807	0.963391
Laughter	0.3812	1.147536	Question ratio	0.0234	0.933465
Overlap	0.0000	46.715341	Ratio of FPSPs	0.0000	0.756782
Bigrams	0.0023	22.650681	Ratio of SPPs	0.0000	0.858746

Table 4: Table of features with corresponding p-values and coefficients in terms of odd ratios, rather than log odd ratios. The model had an accuracy of 0.66 and an F-score of 0.66.

Speaker 1: "I'm sorry if that's how your interactions with friends go 🧐"
Speaker 2: " "Bruh", you're stupid if you think I talked to any of my friends in the manner he posted. I wrote a condensed version of what I did, not verbatim. End result was just people being surprised by how k pop idols look without make up, nothing awkward. I think people on this subreddit project too much when it comes to social awkwardness based on their own inadequacies.
Speaker 1: "Bruh, I don't even like g-dragon but his comment is accurate af in describing how you'd come off as some socially awkward neckbeard if you talked like that to a girl."
Speaker 2: "Good thing I don't actually think that, but the fact that you're even fantasising and joking about it says a lot. Nah, you come off as a loser who doesn't know how to speak to women without scaring them off. That's a sad fact man. Now bye loser "
you."

Figure 1: Balanced conversation where the words that fall into the category interrogations are highlighted with red. In total there are seven words highlighted.

Speaker 1: "lol get fucked. Chinese education is already almost non-existent in Singapore and you want to turn our Chinese syllabus into some half-arsed 'Chinese as a second language'? I'm sorry it offends you that there remain some Chinese people in Singapore who value Chinese as a more than a tool of communication (lol in fact if you really want to participate in meaningful dinner conversation with educated Chinese people, CSL ain't gonna cut it when they name drop historical figures or quote from some poem you've never heard of, much less when idioms are considered assumed knowledge). This is why I wish MOE really just made all the haters drop Chinese or make a CSL course for them and refocus resources into HCL. Besides, in Anglo countries there's generally no separation of 'literature' and 'English' in schools - it's just called 'English'; I don't see why there should be for the Chinese syllabus. "
Speaker 2: "Because Chinese is not the official working language or medium of choice. It's a second language and should be treated as such. I'm sorry that you feel differently but too bad our benevolent dictators have already decided that for you."
Speaker 1: " sorry if I don't have a desire to be a banana "
Speaker 2: "Good for you then, enjoy reading your Chinese classics and consuming Chinese media in your spare time. Just don't fuck it up for the rest of us "
Speaker 1: " lol i'm not even that into Chinese. but anyway, how am i fucking it up for you? you're the ones fucking it up for us. you all should just tell MOE how much you hate Chinese and just drop it."

Figure 2: Unbalanced conversation where the words that fall into the category interrogations are highlighted with red. In total there are seven words highlighted.

3.5 Discussion

Of the 15 LIWC categories, only 5 were statistically significant. It's not surprising that most of these were not significant, since these words can be used in both long comments and short comments. Figure 1 and Figure 2 show 2 conversations, one balanced and one unbalanced, that both have 7 instances of words that fall into the category interrogations. This shows that for this category the amount a certain feature occurs is not correlated with balanced participation. This is the case for most of the LIWC categories.
An interesting one is the category swear words. Reddit being the data set will play a role here, since it is very normal to use swear words in comments. Swearing on informal platforms like Reddit does not necessarily suggest anything negative, but rather it's a way of expression. Would the data set be taken from a social media platform that is more formal, such as a news website with a comment section, swearing would probably not be used in the same manner.

One of the categories that scored very well is conjunctions. It seems that the conversations that have an extremely high number of these words, mostly come from conversations where the first comment is a post rather than an reply comment. On Reddit, people make posts that have a maximum length of 40,000 characters, and others can respond to these posts with comments. The conversations in the data set includes both sequences of just comments, and sequences where a post is the first comment, followed by other comments. These posts can be very long, and since the words in the category conjunctions are generally used often, these posts contain many of these words. Most of the time, the comments following such a post are much shorter than the post itself. This would then be an unbalanced conversation. So, a conversation that has a high number of 'conjuction-words', will most likely be unbalanced, which is probably why this category performs well.

Of the eight other categories, two were not statistically significant, namely laughter and questions. Similar to the other insignificant LIWC categories, laughter could be present in both balanced and unbalanced conversations. This feature only counts the instances that the regular expression recognizes. Moreover, the regular expression used from Bao et al. also counts laughter with the letters 'e' and 'i' instead of 'a', for example 'hehe' or 'hihi'. These forms were not used, in order to keep the expression simple. Other forms of laughter such as the words 'lol' and 'lmao' are not counted towards the total. These words were already part of the category positive emotion, which is why it was decided not to include these in the category for laughter. It would be interesting to see if the addition of these words plus the forms 'hehe' and 'hihi' would result in a better score for this feature.

The feature questions had a p-value a little over 0.05, which is not bad. The value for this feature was the total amount of question marks in the whole conversation, which does not necessarily equal the amount of questions asked. For example, one could write "How are you?" or "How are you????". Both of these sentences contain one question each, but the second sentence contain more question marks. It could be that counting the amount of actual questions rather than the amount of question marks could improve this feature.
The significant features will be used to define three new models, with different combinations of features. The features that did not have a p-value below 0.05 will not be used any further.

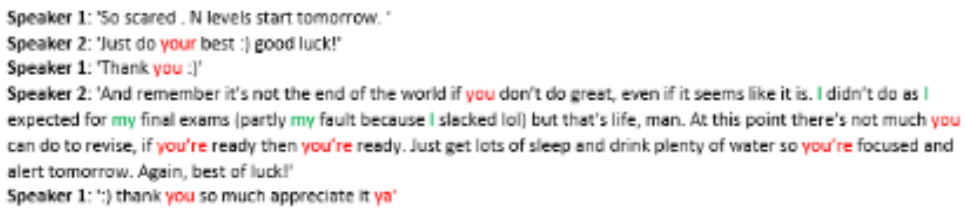

Figure 3: Example of an unbalanced conversation. Second person pronouns (SPPs) are highlighted in red, and first person singular pronouns (FPSPs) are highlighted in green. The ratio of SPPs of speaker 1 to speaker 2 is $\frac{0}{5}$, the ratio of FPSPs of speaker 1 to speaker 2 is $\frac{3}{6}$

4 The Models

4.1 Method

Three different models will be made from the eleven features with a significant p-value.

The first model will contain four features with the best coefficients, namely overlap of words, bigrams, ratio of FPSPs and ratio of SPPs. Four features makes the model quite simple, but since the best four features are chosen, it could still be possible that the model will perform well. This model will be a baseline model that more features could be added to, to increase performance.

The second model will contain all features that are highly significant plus the feature bigrams. The highly significant features are: first person singular pronouns, second person pronouns, conjunctions, length, overlap of words, ratio of FPSPs and ratio of SPPs.
This model is the first model with some features added to it. Adding extra features to the first model may indicate if there is a preference towards either a simple model (only a couple of features) or towards a more complex model (more features).

The features from the first model may be closely related to balanced participation. More specifically, the ratio of certain words spoken by the first speaker to certain words spoken by the second speaker may be highly correlated to the ratio of total words of the first speaker to total words of the second speaker . It seems that these features may have a co-dependency on the proxy, for a ratio of 1 is hardly possible in an extremely unbalanced conversation. Figure 3 shows an unbalanced conversation where speaker 2 uses both first person singular pronouns and second person pronouns more often than speaker 1. For the ratio to become closer to 1, speaker 1 has to add more of these words, which can only happen by adding more sentences. This would in turn make the conversation more balanced. Another possibility is for the second speaker to decrease the usage of these words, but this could decrease the amount of words used, which also makes the conversation more balanced. Either way, how often these words occur may be heavily linked to the total amount of words in the conversation. Predicting balanced participation with features heavily correlated to balanced participation might not be very interesting, and so the third model will not have these features.
Instead, the third model has the remaining features of the eleven with p-values below 0.05, these are: length, first person singular pronouns, second person pronouns, conjugations, positive emotion and perceptual processes.

All three models will be trained on the same train set the feature selection model was trained on. They will then be tested on the test set.

4.2 Results

	Model 1	Model 2	Model 3
Accuracy	0.64	0.68	0.60
Precision	0.64	0.68	0.61
Recall	0.71	0.72	0.62
F-score	0.67	0.70	0.62

Table 5: Accuracy, precision, recall and F-score of all three models. Model 2 scores the highest.

Model 2 scored the best out of the three models, as can be seen in Table 5. This model was an extension of Model 1, and thus adding more features increased performance in all four measures (i.e., accuracy, precision, recall and f-score. Again, the four features from Model 1 have the best coefficients (Table 6). The coefficients of the other features were not as good, which resulted in only a slight increase in performance compared to the performance of Model 1. The feature bigrams performed better in Model 2 than in Model 1, which can be seen from its p-value and coefficient in Table 6 and 7. In Model 1 the coefficient had a value of 9.10, but in Model 2 it had a value of 21.2, almost twice as much. The coefficients of the other features only changed slightly.

Model 3 had the lowest performance, which was expected. This model had all significant features minus the features ratio of questions, bigrams, overlap of words, ratio of FPSPs and ratio of SPPs. Incidentally, these last four features had the best coefficients, and so removing these caused the performance to drop. The remaining features had coefficients close to 1, meaning that the effect they have on the prediction is not much. The features perceptual processes and positive emotions have p-values above 0.05, meaning they are not statistically significantly associated with balanced participation.

4.3 Discussion

Most of the features used in the three models were all significant, which was expected since they were all significant in the Feature Selection Model. However, the features perceptual processes and positive emotion were not significant. This was probably caused by the differences between the validation set, that was used to test the Feature Selection Model on, and the test set, on which the three models were tested.
It is clear from Model 2 that adding more features increases the performance of the model. In this case, the coefficients of the added features were not that high, and so it would be interesting to find more features with better coefficients and see what that does to the performance of the model. The performance would presumably go up with the addition of better features.

Model 1		
Feature	P-value	Coefficient
Bigrams	0.0176	9.104402
Overlap of words	0.0000	44.724235
Ratio of FPSPs	0.0000	0.799528
Ratio of SPPs	0.0000	0.870312

Table 6: The p-values and coefficients of Model 1.

Model 2		
Feature	P-value	Coefficient
Bigrams	0.0024	21.198495
Overlap of words	0.0000	39.383967
Ratio of FPSPs	0.0000	0.754414
Ratio of SPPs	0.0000	0.851793
Length	0.0000	1.066594
First person singular pronouns	0.0002	1.064101
Second person pronouns	0.0000	1.036709
Conjunctions	0.0000	0.967446

Table 7: The p-values and coefficients of Model 2.

Model 3		
Feature	P-value	Coefficient
Length	0.0000	1.062860
First person singular pronouns	0.0003	0.956068
Second person pronouns	0.0000	1.032806
Conjunctions	0.0071	0.987378
Perceptual processes	0.0636	0.983712
Positive emotions	0.6960	0.996988

Table 8: The p-values and coefficients of Model 3.

5 Conclusion

A logistic regression model was built using 23 features, of those, 15 were LIWC categories. Based on the significance of the features to balanced participation, a selection was made to build three new models. In the end, nine of the remaining features were statistically significantly associated with balanced participation, these are the features bigrams, overlap of words, ratio of FPSPs, ratio of SPPs, length, first person singular pronouns, second person pronouns, ratio of questions and conjunctions.
The models using these features were able to predict balanced participation with moderate performance.

The contribution of this thesis is some further understanding as to how balanced participation can be defined and how it exists in online conversations. Balanced participation contributes to the quality of a conversation, and so learning about balanced participation in turn leads to learning about quality of a conversation. Obtaining more knowledge on what makes a conversation "good", may help guide social media platforms to become more civil and enjoyable.

Limitations. The data was split up into three part, that is a train, validation and test set. The total data set consists of conversations from different subreddits, where the conversations were grouped together by these subreddits. The conversations in the train set (partially) belong to different subreddits than the conversations in the validation and test sets. The subreddits differ from each other in terms of topic and types of conversations. Some features will work better on certain subreddits than others, and so a set of features may perform well on the train set, but perform less accurate on the validation or test set. For the features to perform more consistent on all sets, the data could have been randomized. However, this would result in the train set for the Feature Selection Model to be different from the train set for the three final models. But since the features for the final three models were based on the performance of the Feature Selection model, it was decided to keep these train sets the same, rather than randomize them. It would be interesting to see if it would have been better to randomize the data set. Furthermore, the data consists of extreme cases of balanced and unbalanced participation. Conversations that were only slightly unbalanced were not used. Further research could look into the differences between conversations that are not as extreme.

6. References

Alhajji, M., Bass, S., & Dai, T. (2019). Cyberbullying, Mental Health, and Violence in Adolescents and Associations With Sex and Race: Data From the 2015 Youth Risk Behavior Survey. *Global Pediatric Health*, *6*. https://doi.org/10.1177/2333794X19868887

Aroyehun, S. T., & Gelbukh, A. (2018). *Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling*. www.gelbukh.com

Bao, J., Wu, J., Zhang, Y., Chandrasekharan, E., & Jurgens, D. (2021). Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021*, 1134–1145. https://doi.org/10.1145/3442381.3450122

Berry, N., Emsley, R., Lobban, F., & Bucci, S. (2018). Social media and its relationship with mood, self-esteem and paranoia in psychosis. *Acta Psychiatrica Scandinavica*, *138*(6), 558–570. https://doi.org/10.1111/acps.12953

Berryman, C., Ferguson, C. J., & Negy, C. (2018). Social Media Use and Mental Health among Young Adults. *Psychiatric Quarterly*, *89*(2), 307–314. https://doi.org/10.1007/s11126-017-9535-6

Bhat, S. (2016). Effects of Social Media on Mental Health: A Review. *Article in The International Journal of Indian Psychology*. https://doi.org/10.25215/0403.134

Burke, M., & Kraut, R. (2008). Mind your Ps and Qs: The impact of politeness and rudeness in online communities. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 281–284. https://doi.org/10.1145/1460563.1460609

Chang, J. P., Chiam, C., Fu, L., Wang, A. Z., Zhang, J., & Danescu-Niculescu-Mizil, C. (2020). *ConvoKit: A Toolkit for the Analysis of Conversations*. http://arxiv.org/abs/2005.04246

Davidov, D., Tsur, O., & Rappoport, A. (2010). *Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon*. Association for Computational Linguistics. http://tinysong.com/cO6i

Eisenberg, N., & Mussen, P. H. (1977). *The Roots of Prosocial Behavior in Children*.

FitzGerald, N., Carenini, G., Murray, G., & Joty, S. (2011). Exploiting conversational features to detect high-quality blog comments. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6657 LNAI*, 122–127. https://doi.org/10.1007/978-3-642-21043-3_15

Glazzard, J., & Stones, S. (2021). *Social Media and Young People's Mental Health*. www.intechopen.com

Huang, Q., Inkpen, D., Zhang, J., van Bruwaene, D., & Canada, S. (2018). *Cyberbullying Intervention Interface Based on Convolutional Neural Networks*. http://www.bbc.co.uk/news/10302550

Huffman, S., & Ohanian, A. (2005). *Reddit*.

Jiao, Y., Li, C., Wu, F., & Mei, Q. (2018). Find the conversation killers: A predictive study of thread-ending posts. *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, 1145–1154. https://doi.org/10.1145/3178876.3186013

Jurafsky, D., Leskovec, J., Danescu-Niculescu-Mizil, C., Sudhof, M., & Potts, C. (2014). *A computational approach to politeness with application to social factors*. http://en.wikipedia.org/wiki/

Kemp, S. (2020). *October Global Statshot Report*.

Makri, S., & Turner, S. (2020). "I can't express my thanks enough": The "gratitude cycle" in online communities. *Journal of the Association for Information Science and Technology*, *71*(5), 503–515. https://doi.org/10.1002/asi.24257

Martela, F., & Ryan, R. M. (2016). Prosocial behavior increases well-being and vitality even without contact with the beneficiary: Causal and behavioral evidence. *Motivation and Emotion*, *40*(3), 351–357. https://doi.org/10.1007/s11031-016-9552-z

Martin, T. (2018). *Woebot: A smart, accessible mental healthcare solution*.

Napoles, C., Pappu, A., & Grammarly, J. T. (2017). *Automatically Identifying Good Conversations Online (Yes, They Do Exist!)*. www.aaai.org

Napoles, C., Tetreault, J., Rosato, E., Provenzale, B., & Pappu, A. (2017). *Finding Good Conversations Online: The Yahoo News Annotated Comments Corpus*.

Peterson, M. (2021). *How Much Time Do People Spend On Social Media In 2021*.

Radev, D. R., Hassan, A., Qazvinian, V., & Radev, D. (2010). *What's with the Attitude? Identifying Sentences with Attitude in Online Discussions* (Issue 11). Association for Computational Linguistics. https://www.researchgate.net/publication/221012779

See, A., Roller, S., Kiela, D., & Weston, J. (2019). What makes a good conversation? How controllable attributes affect human judgments. *NAACL-HLT*.

Singh, R., Subramani, S., Du, J., Zhang, Y., Wang, H., Ahmed, K., & Chen, Z. (2020). Deep learning for multi-class antisocial behavior identification from Twitter. *IEEE Access*, *8*, 194027–194044. https://doi.org/10.1109/ACCESS.2020.3030621

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. In *Journal of Language and Social Psychology* (Vol. 29, Issue 1, pp. 24–54). https://doi.org/10.1177/0261927X09351676

Wadden, D., August, T., Li, Q., & Althoff, T. (2020). *The Effect of Moderation on Online Mental Health Conversations*. http://arxiv.org/abs/2005.09225