Dialogue Classifiers to Improve Access to Mental Health Services

Problem presented by

Joseph Connor (NHS Digital)

ESGI138 was jointly hosted by

The University of Bath

The University of Bristol

with additional financial support from

Bath Institute for Mathematical Innovation (IMI)

The Engineering and Physical Sciences Research Council (EPSRC)

GW4 Alliance

The Infrastructure Industry Innovation Platform (I3P)

Industrially Focused Mathematical Modelling CDT (InFoMM))

Innovate UK's Knowledge Transfer Network (KTN)

Report Authors and Contributors

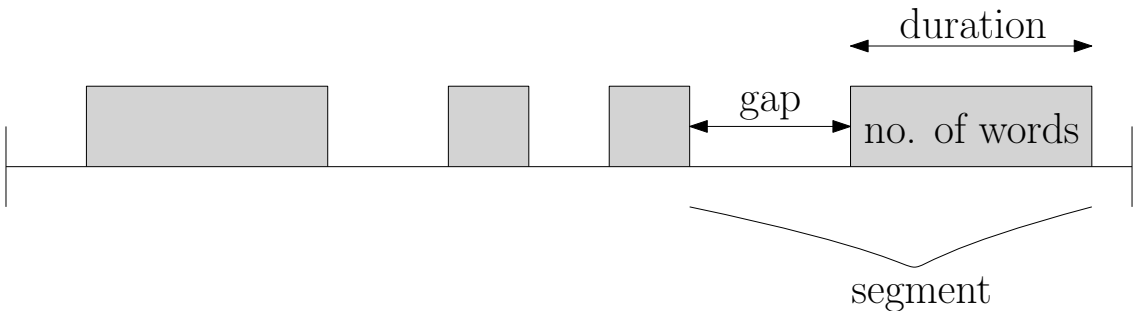
Facilitator:
Lorna Wilson [1],

Contributors:
Radu Cimpeanu[2], Matthew Moore[2], Oliver Sheridan-Methven[2], Lingyi Yang[2], Zhen Shao[2], Melanie Beckerleg[2], Ana Osojnik[2], Yuanwei Xu[3], Tsung Fei Khang[4], Choung Min Ng[4], Hamzah Nor Aishah[4]

Acknowledgements
This project has been supported by GW4 data science seed corn funding.

Contents

[1] University of Bath, UK
[2] University of Oxford, UK
[3] Imperial College London, UK
[4] University of Malaya, Malaysia



[Ana: it would be good to add this figure somewhere in the section which describes different features.]

1 Executive Summary

2 Problem description

Millions of people call UK healthcare service providers in distress because of mental health issues. Unless a call handler, who may not have any background in mental health support, can deescalate the conversation to the point, where the caller is prepared to accept advice, they will most likely present themselves at A&E which may not be the best place for them. Therefore, providing call handlers feedback on whether the dialogue is going well or badly, perhaps via a traffic light system, could prove valuable, as it may indicate if the conversation is going in the direction of deescalation, and the caller is more prepared to take advice. Clinicians report that when callers feel less anxious, they are more likely to listen to the advice given, and follow it. A simple method for classifying the distress level of a call as 'good' or 'bad' could be used to provide real-time call handler feedback.

2.1 Is it possible to create a classifier for 'good' and 'bad' dialogues to improve access to mental health support?

While this may seem like a typical NLP/DNN problem, NHS requires a different approach for a number of reasons:

- Data protection and GDRP makes the NLP/DNN approach difficult/impossible.
- A method that can be explained in simple terms to clinicians and call handlers is more likely to be trusted and used by them. For example, a highly accurate black-box method is not as useful as a more interpretable method with slightly lower accuracy.
- Analysis needs to be possible, cheaply and close to real-time. It should be feasible to integrate it with the existing pipeline, such as real-time speech recognition.

2.2 Data availability

Highly personal data, call recordings and their subsequent outcomes cannot be made available to the Study Group. Clinicians report that, in practice, a conversation can be described as less distressed when it resembles friendly chatting, where participants are listening and responding to each other, whereas a conversation with a highly distressed individual will sound more like a difficult or confrontational political interview. [Ana: I think this is wrong. See corrected text in green.] ~~The NHS has access to transcripts, provided by their partners, for both types of correspondence.~~ Therefore, the NHS has acquired audio recordings of conversations, which are believed to be to some extent resemblant of the two types of correspondence. Each of the dialogues was evaluated by

a group of clinicians as a 'good' or a 'bad' conversation. The idea is to use these conversations as examples of 'good'/non-distressed (93 transcripts) and 'bad'/distressed (28 transcripts) NHS 111 calls for the purpose of developing initial proof-of-concept methods. Every conversation in this example dataset was machine transcribed (with a degree of inaccuracy), which, in addition to the words said, also included recording the time passed between any two consecutive words, and the speaker. Attempting to circumvent the issues around data protection and limited accessibility of NHS call transcripts in real life, the example transcripts were converted into numbered words, with time from and to when the word was said, gaps between words, and speaker identifier. We aim to develop a classifier for a dataset of such a form (see Figure 1).

[Ana: I do not understand this. Needs to be made clearer.] There are also 'sentiment' values. These are not considered to be a reliable measure, but they may be useful for tracking changing sentiment within a conversation.

Conversation	1	1	1	1
Word	TimeFrom	TimeTo	Speaker	Gap between speakers
1	2.23	2.36	3	0
2	2.36	2.76	3	0
3	2.76	3.28	3	0

Figure 1: The format of data available for one example conversation

2.3 Mathematical questions

During the Study Group, our aim was to answer the following questions:

- Can features such as gaps in speech and/or turn-taking be used to identify or predict whether an entire conversation is 'good' or 'bad'?
- How could this be applied to the data in a near-real-time feed?
- What models/techniques should be considered that do not require the creation of a transcript from a live, two person, single file audio, telephone call?
- Can the features and methods considered also be used to identify or predict when a conversation is changing from 'good' to 'bad' or vice versa?

3 Data exploration

3.1 Assessing the quality of the data

With all data driven problems, a useful first step is to understand how the data were gathered and to perform data cleaning if possible/necessary. In the case of this Study Group, because of privacy issues, actual NHS helpline phone calls were not available for analysis. The alternative suggested by NHS is to test their hypothesis of turn-taking on audio clips from the Andrew Marr show and the BBC Listening Project. It is supposed that Andrew Marr show, which hosts political interviews, is negative since the speakers are more confrontational so these clips are classed as 'bad'. On the other hand, the BBC listening project are typically interviews with two or more participants who already have rapport between them, for example colleagues, friends or families. Because the interviewees can take their time to formulate their answers, these conversation are more productive and thus classed as 'good'. We have 28 samples from the Andrew Marr show and 93 samples from the BBC listening project. One thing to note here is that the sample sizes are inbalanced and indicates reporting the absolute accuracy of classification may be misleading (for example just classfying all clips as 'good' will achieve an accuracy of 76.9 % but it is a very poor method), this is important when we progress to comparing our methods.

From the audio clips of the two sources, information about which person is speaking, when the words are spoken and the pauses between one speaker and the next are extracted. This process is performed by five different people. These five people may have used different laptops with different settings and software to record/convert the audio into the data that we have. This means that there could be some systematic inconsistencies.

Very soon into the data exploration, some problems with the data are seen. Mislabelling of the headings resulted in mismatched 'conversation' numbers so rather than 93 clips from BBC listening Project, we instead have only 88 samples of 'good'. With further analysis of summary statistics, the data appears to be repeated on at least two occasions. We see that the data for 'good-50' and 'good-51' are identical, so are 'good-53' and 'good-54'. Note that we refer to the original labels which are mismatched so that it can be easily checked in the dataset. Again this is very puzzling and would be best avoided/ easily checked by keeping a clear record of where the audio clips came from.

We note that one clip of the 'bad' conversations appears to be a monologue, this is thought to be rather strange given the hypothesis that we are testing. Hence some audio clips were physically listened to along with reading the transcript which corresponds to the words and times that these words were said. For this particular example of 'bad' which appeared as a monologue, the original audio file as well as the transcript cannot be located. It is not known exactly which interview on the Andrew Marr show it

corresponds to and no further information can be found.

However, it is still worth checking some of the other audio files and transcripts which we can access for better understanding of the data. By listening to these, it can be observed that the data we have are both unreliable and noisy. Many of the one-word 'dialogue' are simply artefacts and the program is not picking up the real one-word response (such as responses like 'yeah'). This means that some important indicators may have been missed.

Exchanges between the speakers can be inaccurate, for example in good-10, there was more than 10 exchanges between two people but this was simply recorded down as a chunk of over 600 words by one person. This does not just appear in the cases where the voices are similar but also when there are male and females voices. Perhaps another question for future voice analysis is which audio-to-text programme is most reliable in distinguishing between different speakers.

In other cases, changes in the tone or modulation results the algorithm in labelling that the speaker has changed even though it hasn't. There are also other seemingly random cases where a change in the speakers was recorded which cannot seem to be explained by the audio files or transcript.

Similarly, the algorithm often classifies the speach into that of three different speakers, even though in reality there are only two people involved in the conversation.

Such evidence raises the question of exactly how accurate the data are for the purpose of our investigation, since segmention of speech by speaker is not reliable and the gap between speakers may simply be one person pausing to speak. Therefore we should proceed with caution. All audio was subjected to a similar process, therefore one might assume there may something systematic in the inaccuracies of the data. We must remember this problem when it comes to drawing conclusions.

If feasible, the gathering of the data should be redone before verifying the findings of this report. Although parallelisation of the work is more efficient, for purposes of testing techiques, it is better to have the data collected by one laptop with the settings kept constant. Another recommendation is keep a safe record of the original source so that it could be referred upon if needed, for example when checking anomalies. Once collected, a quick check with the audio files can be performed to see if the quality of the data have improved since this initial work at the study group.

3.2 Basic analysis of provided datasets

The provided "good" and "bad" datasets allow the easy extraction of conversational metrics that would assist in the classification of the respective dialogue. As an initial attempt and guided by previous efforts, we consider several so-called standard features that naturally arise from the dialogue format, namely:

1. Number of words for the respective speaker in one continuous segment.
2. Duration. This pertains to the entire active speech period for a given speaker, measuring the time from the start of the first word to the end of the last word of the respective person before being interrupted.
3. Inter-speaker gap. Measures the time between the change in speakers, namely from the end of the last word of the previous speaker, to the start of the first word of the following speaker.

These three features characterise each individual conversational segment, which is preceded or followed by similar groupings.

We begin with a basic statistical analysis of the three metrics listed previously, with a focus on both means and variability (expressed through standard deviations) of the respective datasets. Figures 2 and 3 provide an illustration of our findings.

Each symbol in these two figures pertains to a selected conversation, with the dashed lines indicating the mean of the entire dataset in each individual subfigure. We note that in the interest of aligning this investigation with the subsequent probabilistic and machine learning approaches, we have discarded the first third of conversations as supplied in the Excel spreadsheets, restricting ourselves to the final 20 "bad" conversations and 57 "good" conversations.

In the "good" dataset, the mean for the number of words is calculated to be at approximately 84, with a standard deviation (st. dev.) of 115.6. The duration (in seconds) for each such segment has an average of 27.95 s with a st. dev. of 38.9 s, while the gaps between speakers average at 0.565 s with a st. dev. of 0.594 s. Even before comparing to the other set of conversations, a key feature of the data emerges in the form of the very high variability present throughout the family of basic features considered. The reliability of the conclusions formulated strictly on the basis of the datasets is hence called into question.

A similar trend is observed for the "bad" dataset, with standard deviations revealed to be roughly as high as the means of the quantities considered yet again. There are noticeable differences in terms of the selected features, with significantly less words 38.52 with 36.69 st. dev., a roughly fourfold decrease in the mean - 11.06 s - with 10.82 st. dev. for the duration and 0.343 s with 0.39 st. dev. for the gaps. All these are indicative of a more abrupt dialogue, with more frequent and more sudden changes in speaker. However ultimately the suitability of the monitored quantities appears to have a strong dependence on the nature of the data (see comments in the previous subsection) and its extraction technique as opposed to inherent features of good/bad dialogues. With relatively little information (very few data points) in the context of learning algorithms, the key conclusion from these preliminary statistics is that any successful algorithm should be built on a more realistic (and richer) data framework.

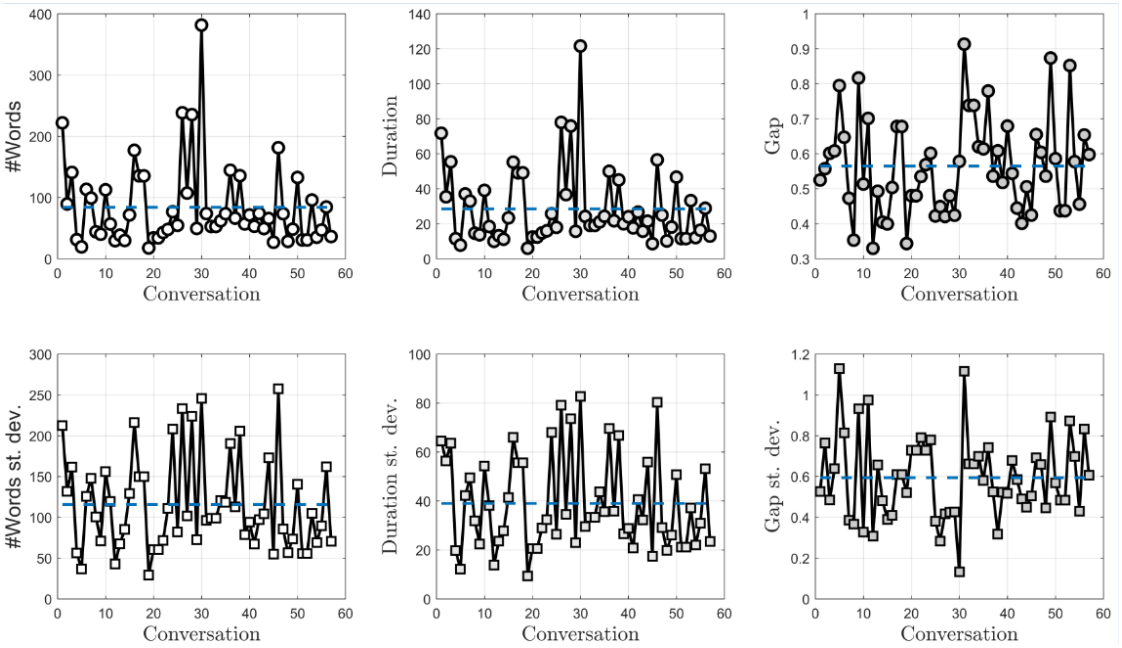


Figure 2: Means (top row) and standard deviations (bottom row) for the primary features in the good dataset - number of words (left), duration of speaker block (center) and gap size (right). The dashed lines indicate means across the entire dataset depicted in each subfigure.

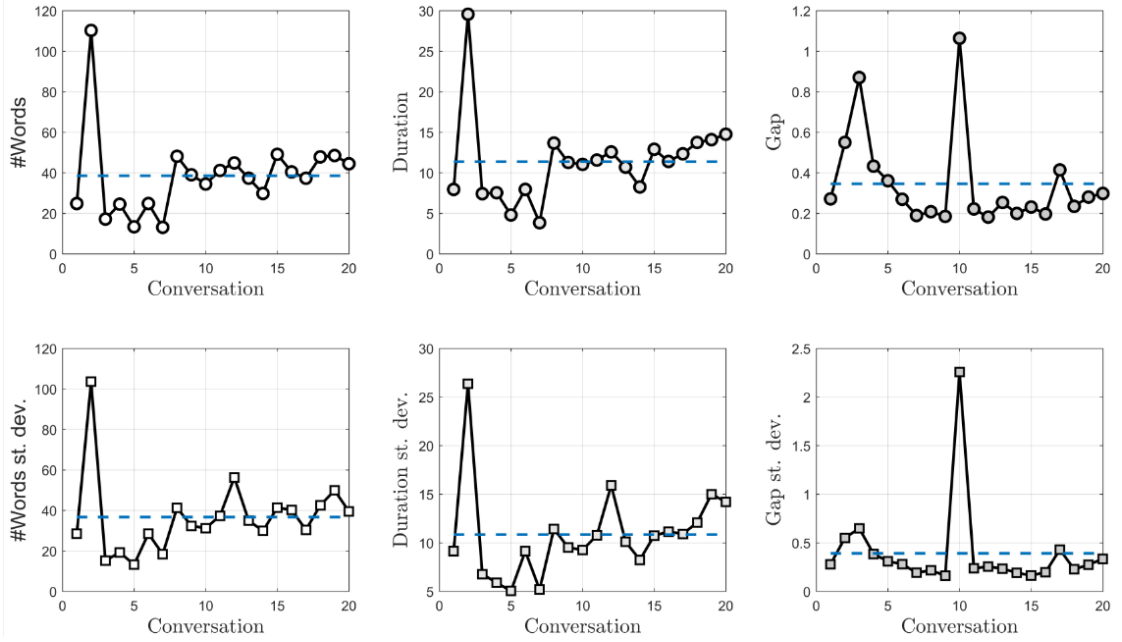


Figure 3: Means (top row) and standard deviations (bottom row) for the primary features in the bad dataset - number of words (left), duration of speaker block (center) and gap size (right). The dashed lines indicate means across the entire dataset depicted in each subfigure.

3.3 Further metrics

Apart from the basic statistics on already known quantities, efforts have been made to detect new and perhaps better suited features for dialogue classification. Even if not entirely successful on the particular present dataset, these are intended as alternative quantities to consider in the future as the platform gains maturity.

3.3.1 Interruptive behaviour

One aspect of particular importance in this application is the ability to calculate and adapt methodologies on the fly, based on the progress in the conversation itself as opposed to judging the outcome at the end of each call. Thus integrating features which respond at each turn (henceforth marked as t) is a desirable addition. One such indicator, taking into account the *evolution* of the gaps within each conversation is an "interruptor function" f_i, which could be defined as:

$$f_i(t) = \begin{cases} f_i(t-1) - 1 & \text{if } g(t) < g(t-1) \\ f_i(t-1) + 1 & \text{if } g(t) \geq g(t-1) \end{cases}$$

Incrementing a function by plus or minus 1 depending on whether the current gap $g(t)$ is larger or smaller as the conversation progresses may alert as to whether the speakers have a tendency to interrupt themselves and engage in a more aggressive dialogue. We anticipate "bad" conversations to have a lower f_i score and Fig. 4 does indeed confirm this tendency, with an average score of -3.65 for the bad conversations (above) to the -1.1 observed in the good conversations (below). Naturally, several layers of complexity can be added to such constructs, with the ± 1 modification being replaced by the gap differential $g(t) - g(t-1)$ itself, perhaps weighted by the average gap size as given by the historical dataset. Such upgrades have been attempted, with the overall conclusion and variation being maintained irrespective of the particular choice. The bottom line of such methods is that they offer another more adaptive perspective in the evolving conversation and such a score calculated on the fly may be used as another classifier when a certain threshold is crossed.

Figure 4 also shows so-called streaks in gap evolution, namely the longest number of segments for which monotonical behaviour is observed. For example, a gap increase streak of four indicates that over four exchanges between the speaker, the gap has grown every single time as compared to the previous gap and at the fifth exchange a decrease was observed. Conversely, a gap decrease streak shows that the gaps become progressively smaller, which may be an indication of sustained aggression by all conversational partners and indeed an increase in the intensity of the dialgoue. We expect this to be more typical of the "bad" conversations. No significant trends have been recovered from the test dataset however. Again, one may wish to consider variations of such metrics, for example adding a buffer for what constitutes a genuine variation (as opposed to small machine detection/precision level changes).

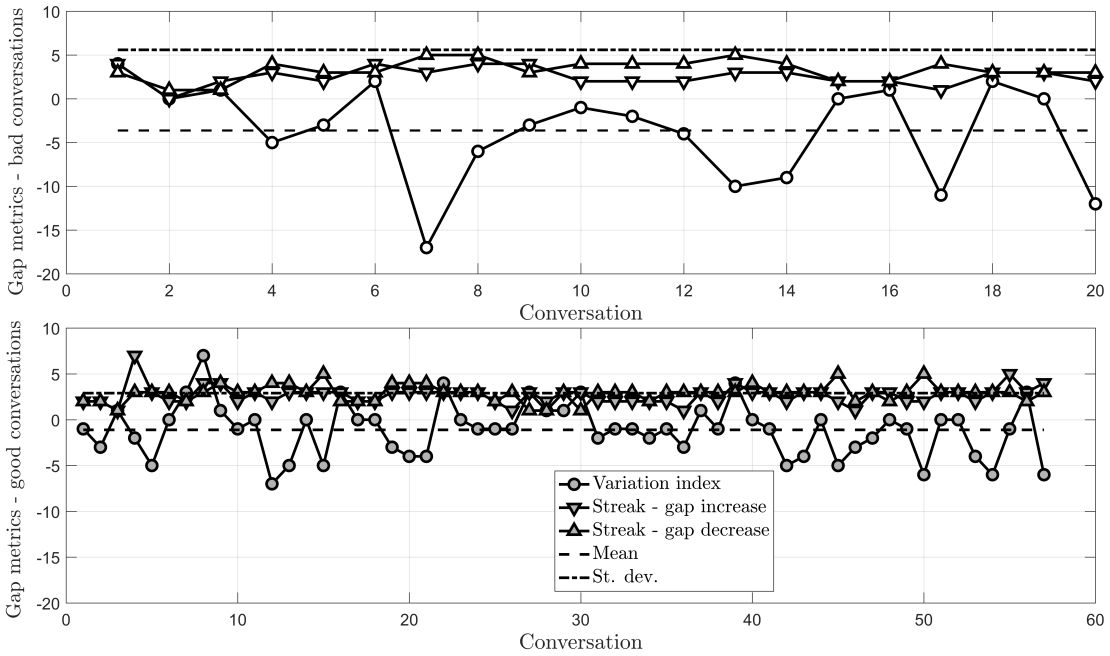


Figure 4: Metrics on the gap evolution within each conversation, detailing on the variation of the so-called interruptor function, as well as trends in terms of increasing or decreasing gap sizes for both bad (above) and good (below) conversation datasets.

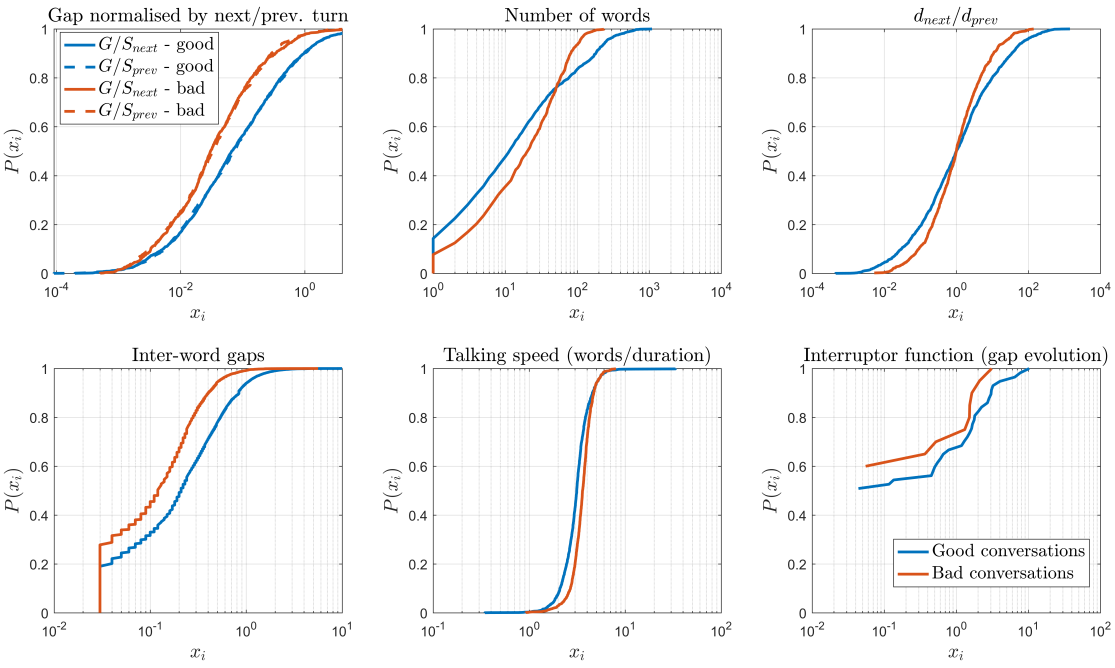


Figure 5: Cumulative distribution functions for both good and bad datasets expanding on the most promising features based on the exploratory data analysis.

The key outcome of this exercise is to address the question of which features are suitable classifiers from a purely statistical standpoint. After analysing the constructed datasets, cumulative distribution functions (cdf's) can be used as both a visual indica-

tor and as building blocks for quantitative assessments (e.g. the Kolmogorov-Smirnov test) that follow. The summary of the most promising features is presented in Fig 5, in which all "bad" conversations are shown in red, while the "good" conversation data is presented in blue. The larger the absolute difference in the two curves (irrespective of the sign of the difference), the better the respective feature is expected to perform as a classifier. We notice that features such as the gap normalised by the previous or next turn, the number of words (the most revealing of the basic features studied initially when used in isolation), as well as the presented interruptor function provide good candidates. The duration of the next versus the preceding segment, the gaps between words themselves and the talking speed (number of words divided by duration) appear to be at first glance less promising classifiers. Some of these are simply due to the statistical similarity of the datasets considered, while others have systematic issues in the dataset given by the dialogue format or by the engine translating the speech to text. This is the case of the inter-word gaps, in which quite often translation errors lead to artificial changes of speaker and fewer than expected data points in general due to the difficulty in measuring pauses between words.

We stress yet again that many of these conclusions should be assimilated noting the high dependence to a specific small dataset and may prove to be fruitful alternatives as the application-specific data becomes available.

In addition to the three 'standard features' outlined at the start of this section and the possible composite features and functions mentioned in the previous paragraphs, we also looked at some further features of dialogue that may provide insight into how well a conversation is progressing, which we discuss here.

3.3.2 Speed of speech

At the end of the previous section, we briefly mentioned the 'speed of speech' as a possible metric for characterising good or bad conversations, where speed of speech is defined by

$$\text{Speed of speech} = \frac{\text{Number of words in a segment}}{\text{Duration of segment}}.$$

We would expect that the speed of speech would on the whole be slower in good conversations, where the individual speakers take longer time to digest what their partner is saying before responding. Conversely, we would expect bad conversations to be characterised by interruptions and pre-prepared segments from each speaker delivered irrespective of what the previous speaker said.

The results are depicted in figure 6. We have depicted the results for all conversations in the respective datasets together, rather than investigating each dialogue on a case-by-case basis to investigate the overall trends. Although there is a large amount of noise in the data — particularly for the good dialogues — we do see evidence of the expected behaviour, with the mean of the speed of speech slightly higher for the bad

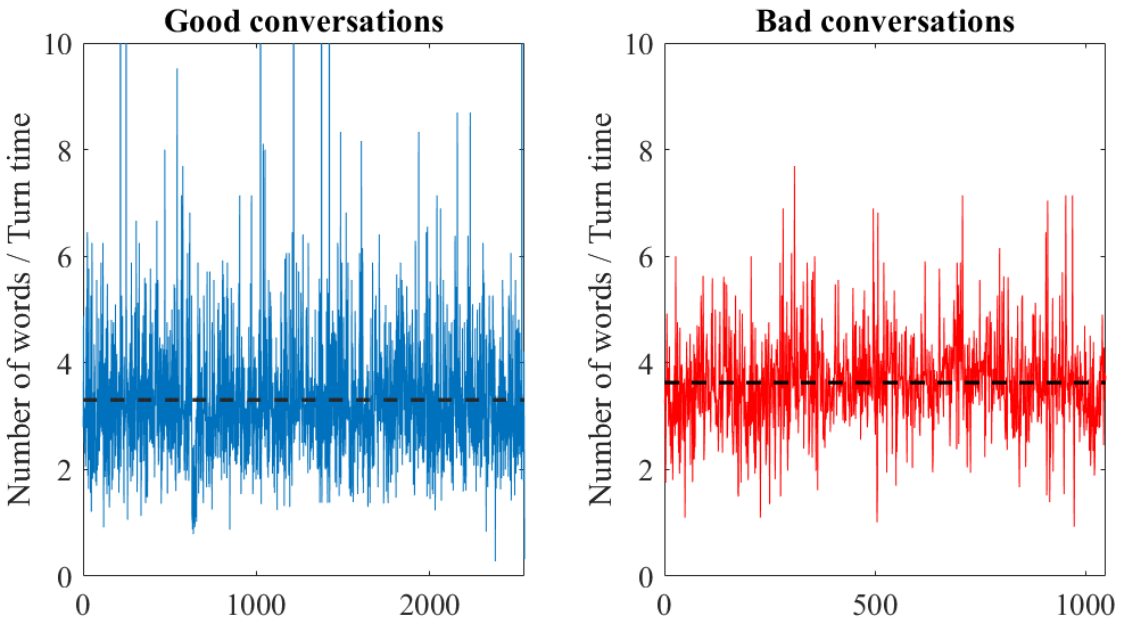


Figure 6: The speed of speech as defined by (3.3.2) for all conversations (excluding test data) in each of the good and bad datasets. In each of the subplots, the dashed line indicates the mean speed of speech.

conversations than the good. Specifically, for the good conversations the mean number of words per segment is 3.31, while for the bad conversations it is 3.63.

Therefore while it is quite difficult to make firm conclusions based on the relatively small difference between the means and the large amount of noise in the data, it is reasonable to expect that with a better dataset the speed of speech may be a good metric to distinguish between good and bad conversations. We will now utilise speed of speech to investigate a further property of 'good' conversations.

3.3.3 Mimicry

A further feature of 'good' conversations is that speakers tend to speak with similar cadence or speed. This is a result of the speakers feeling more comfortable as the dialogue progresses and reacting the speed of speech of the other speaker(s).

In order to visualise how the speed of speech changes as a conversation progresses, we plot the speed of segment $i+1$ against segment i in figs. 9 and 10 for various conversations within each dataset. Within each subplot, the crosses get darker as the conversation progresses. Therefore, we see evidence of mimicry if the crosses start to cluster towards the same value as the crosses become darker.

We analyse half of the good conversations (excluding the test datasets) in figures 7–9. As the bad dataset is smaller, we have only analysed nine of the conversations, as plotted

in figure 10. Throughout, we encounter the aforementioned issues with the provided datasets: very short dialogues where it is difficult to gauge any meaningful changes within the short dialogue duration; dialogues dominated by short responses from one speaker; difficulties in ascertaining when there is a change of speaker and/or when there are more than two speakers within a conversation. These issues notwithstanding, we can take away some insights from these figures.

If we restrict ourselves to dialogues of a reasonable lengths, we do see evidence of clustering as the crosses darken in the good datasets in several, namely in dialogues 34, 44, 50, 68, 72, 78 and 82. On the other hand, for dialogues 42 and 84 for example, there is still a large amount of scatter as the conversation progresses, indicating no sign of mimicry in these dialogues. In contrast, although the sample size is smaller, we do not see evidence of clustering in the bad dialogues: indeed dialogues 11, 13, 17, 21 and 25 appear to show a large amount of scatter as the conversation progresses, suggesting that the lack of rapport between the speakers is not leading to mimicry in these conversations.

With a more realistic dataset it would be possible to investigate this in more detail, but it is encouraging to see some evidence of expected behaviours in the conversations.

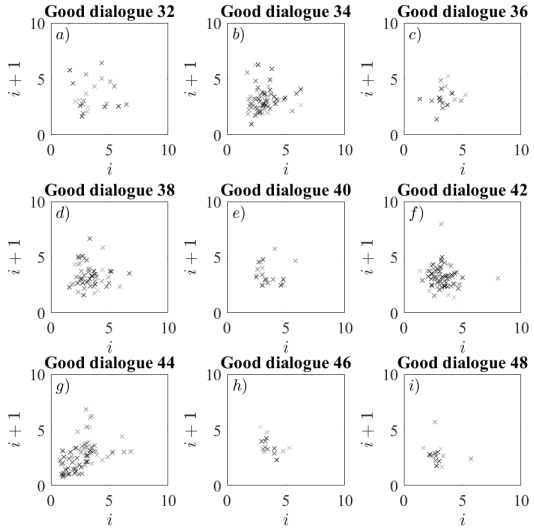


Figure 7: The speed of speech of segment i against segment $i+1$ for nine of the conversations in good dataset: a) 32; b) 34; c) 36; d) 38; e) 40; f) 42; g) 44; h) 46; i) 48. Within each figure, the cross becomes darker as the conversation progresses. For evidence of mimicry, we would expect to see a pattern of the crosses clustering as they become darker.

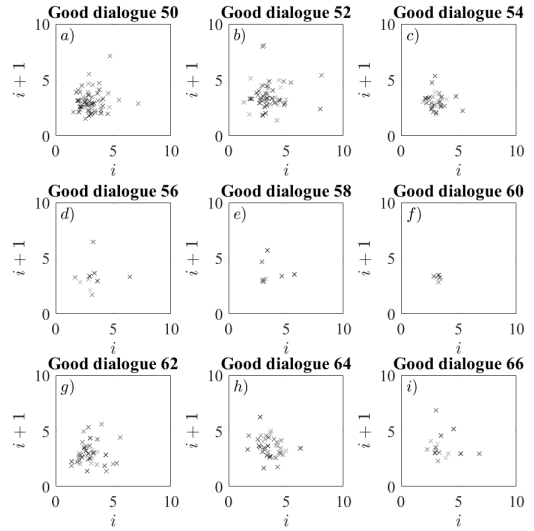


Figure 8: The speed of speech of segment i against segment $i+1$ for nine of the conversations in good dataset: a) 50; b) 52; c) 54; d) 56; e) 58; f) 60; g) 62; h) 64; i) 66. Within each figure, the cross becomes darker as the conversation progresses. For evidence of mimicry, we would expect to see a pattern of the crosses clustering as they become darker.

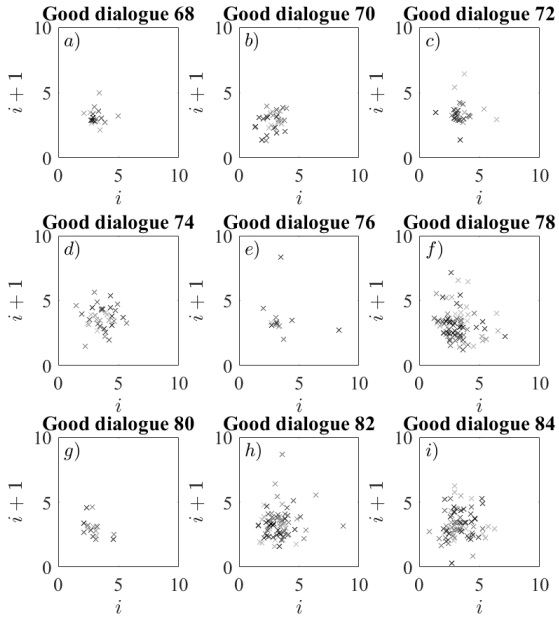


Figure 9: The speed of speech of segment i against segment $i + 1$ for nine of the conversations in good dataset: a) 68; b) 70; c) 72; d) 74; e) 76; f) 78; g) 80; h) 82; i) 84. Within each figure, the cross becomes darker as the conversation progresses. For evidence of mimicry, we would expect to see a pattern of the crosses clustering as they become darker.

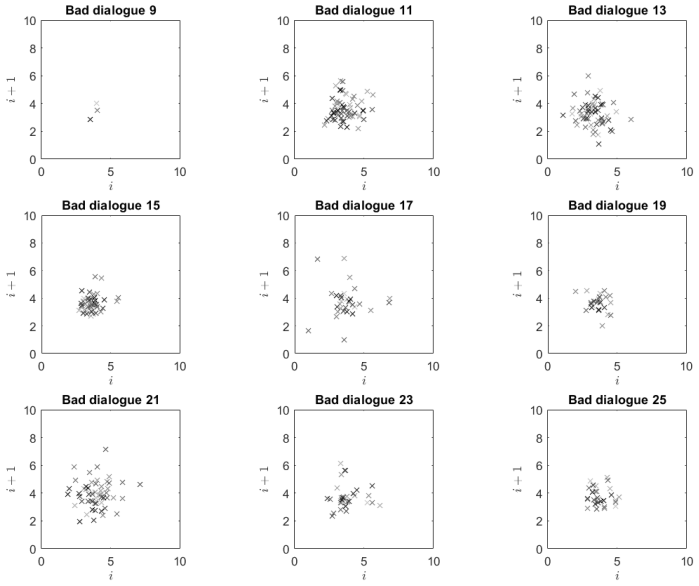


Figure 10: The speed of speech of segment i against segment $i + 1$ for nine of the conversations in bad dataset: a) 9; b) 11; c) 13; d) 15; e) 17; f) 19; g) 21; h) 23; i) 25. Within each figure, the cross becomes darker as the conversation progresses. For evidence of mimicry, we would expect to see a pattern of the crosses clustering as they become darker.

4 Bayesian updating scheme

A natural approach to take when tackling this problem is to turn to Bayesian statistics. We can see by the description of the problem that call handlers update their estimation about whether the conversation is going well or badly, based on the information that becomes available to them as the conversation progresses. Immediately we notice that this fits within a Bayesian updating framework for hypothesis testing over live-feeding data. (See Grindrod [1, Chapter 5] for a more detailed discussion of the underlying mathematics presented in this section).

If we denote $\mathbb{P}(A)$ as the probability of an event A happening, $\mathbb{P}(A \cap B)$ as the probability of both A and B happening, and $\mathbb{P}(A \mid B)$ as the probability of A happening given that event B has occurred, then we recall Bayes' theorem [1, pages 219–223]

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1}$$

For convenience we also define the odds $\mathbb{O}(A)$ of an event A happening as

$$\mathbb{O}(A) \coloneqq \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)}. \tag{2}$$

In the setting of the call sentiment classification, we want to assess the odds that the conversation is going well or going badly. We formulate such a categorisation in terms of two hypotheses (events), where H_0 is the null hypothesis that the conversation is a 'bad', and H_1 is the alternative hypothesis that the conversation is a 'good'. Note that the two hypotheses are mutually exclusive, i.e. $H_0 = H_1^c$.

If after a duration of conversation we have the set D of observations, and we observe a new observation d, then we can transform eq. (1) into the Bayesian updating scheme

$$\underbrace{\mathbb{O}(H_0 \mid \{D \cup d\})}_{\text{posterior}} = \underbrace{\frac{\mathbb{P}(d \mid H_0)}{\mathbb{P}(d \mid H_1)}}_{\text{model}} \underbrace{\mathbb{O}(H_0 \mid D)}_{\text{prior}}. \tag{3}$$

It is worth taking a moment to explain eq. (3), which comprises three terms: the prior odds, the model, and the posterior odds. The prior odds are easily interpreted as the odds of H_0 being true given all the data D, and hence they correspond to the odds of the conversation going badly given all the information that has been exchanged between the conservation participants. When a new exchange d has occurred in the conversation, the model is the ratio between the likelihood of the new observation if the conversation is indeed 'bad', and the likelihood under the assumption that the conversation is 'good'. The posterior are the new updated odds given all the information up to this point in time. We see that we can apply eq. (3) iteratively as new pieces of information become available to compute live updates of the odds that the conversation is going well/badly.

4.1 Creating the models

There are several ways to form the models used in eq. (3), and the simplest approach is to try and extract these probabilities from the distribution of training data. To illustrate this procedure, we consider the histogram of the duration of speech segments gathered across all conversations in the training dataset, shown in Figure 11.

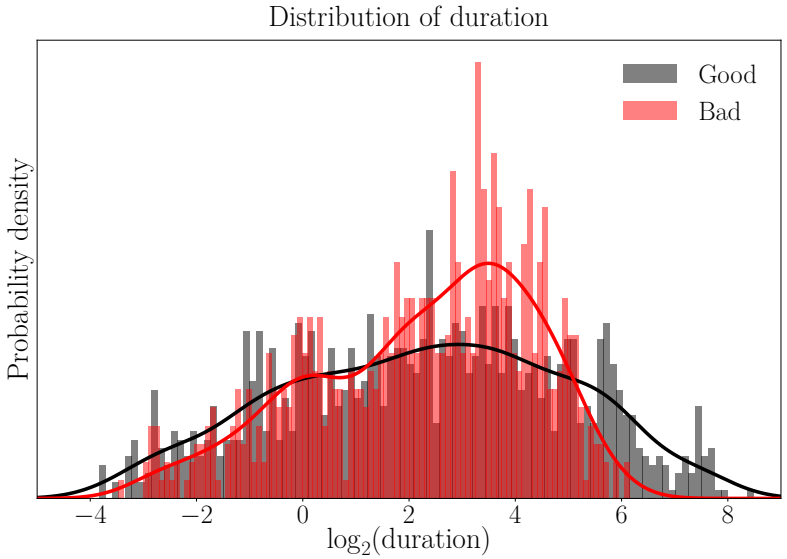


Figure 11: The empirical estimates of the probability density function for the duration of speech segments. The histogram shows the distribution of the values, and the solid lines depict the estimates of the probability density functions obtained by using a non-parametric Gaussian kernel density estimate.

We can see from the histograms in Figure 11 that the durations (specifically the logarithm of the duration) follow similar distributions, but the 'good' conversations appear to have a larger variance. These histograms though are not particularly smooth due the limited size of the training data, and hence we form a smoothed approximation using a Gaussian kernel density estimate.

To demonstrate how we might update the current odds held at a point in time, suppose we receive a new observation, i.e. a segment of speech, and at the next time point we measure its duration. When the duration corresponds to a value in Figure 11 where the black curve is above the red one, i.e. the observation is more likely to be from a 'good' conversation than a 'bad' one, then we calculate the ratio of probability densities on the two curves at this value, and scale the odds by the computed ratio according to eq. (3). Alternatively, for the given duration, if the red curve is above the black one, then we decrease the odds with such an update, whereas if the curves intersect (or are very close), then the odds do not change (substantially).

We can update the odds in this way for any features we have available, such as the segment's duration, preceding gap length, or the number of words said in a segment.

It is important to note that features such as the duration of a segment of speech are highly correlated (effectively proportional) to the number of words said in it, and hence this characteristic of a segment will have twice the influence on the change of odds than the gap length.

4.1.1 Optimal feature weighting

It is a separate challenge to compute the optimal weighting of the features and decide how heavily each should influence the odds. This was briefly explored for the scheme that only used the three features mentioned, although this quickly became a challenging optimisation problem we had to tackle using brute force derivative-free optimisers. We mention this in passing as this would remain an important step for fine tuning any implementation, and is a separate challenge in itself.

4.2 Choosing the prior

Choosing the prior term in eq. (3) is very much at the discretion of the modeller. In the absence of any data ($D = \emptyset$), a good estimate is to set $\mathbb{O}(H_0 \mid \emptyset) = \mathbb{O}(H_1 \mid \emptyset) = 1$. However, for a slightly more informed prior we can turn to the ratio between 'good' conversations and 'bad' ones as seen in the training data, which were in a relation of about 2:1 respectively, thus giving $\mathbb{O}(H_0 \mid \emptyset) = 1$, $\mathbb{O}(H_1 \mid \emptyset) = 2$. However, we will see from the results that this has little impact.

4.3 Running odds

Using the updating scheme from eq. (3), we can see how the odds evolve as the conversations progresses. We plot the running odds as computed for the testing set in Figure 12. From this plot we can see that the scheme separates out well the 'good' from the 'bad' conversations as the conversations progress.

If the reader is alarmed at seeing such astronomically large and small odds, this is a by-product of all features being given an equal weight of 1. If we instead allowed all features to have equal weights, but constrained the sum of all weights to 1 (a useful constraint for an optimisation routine), then for three equally weighted features each feature would have a weight of $\frac{1}{3}$. Therefore, changing the weight from $1 \to \frac{1}{3}$ would have the effect of reducing the exponents of the odds in Figure 12 by a factor of 3.

To assure ourselves that such a scheme translates well to new and unseen data, we plot the same results as computed for the validation set, and also for a real-life clinical conversation, shown in Figures 13 and 14 respectively. We can see that in validation we achieved a good (but not perfect) separation of the 'good' and 'bad' conversations. Furthermore, we also correctly classified a real world clinical conversation using this

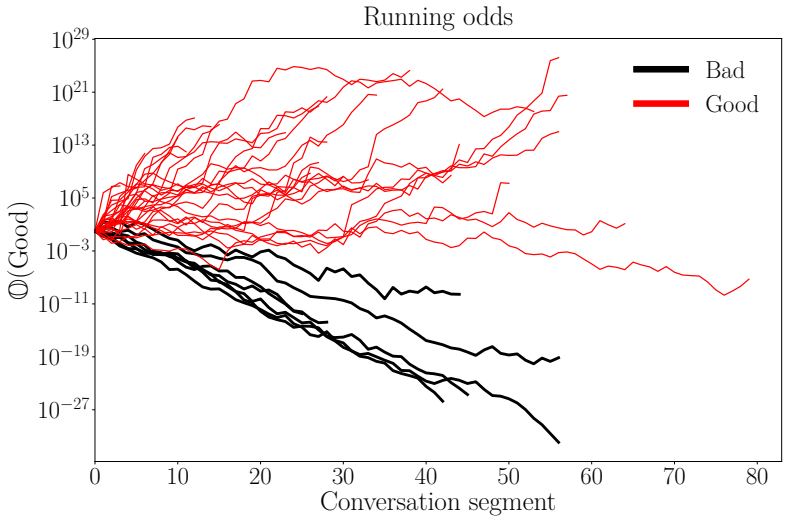


Figure 12: The running odds from eq. (3) on the testing set using a model derived from Gaussian kernel density estimates for the logarithm of the values of the duration, gap length, and number of words in a segment.

scheme. However, this was only a single correct classification so **the reader should not attribute significance to this one-off success until verified on a statistically significant real-life validation set, which is relevant for the application and consists of a large number samples.**

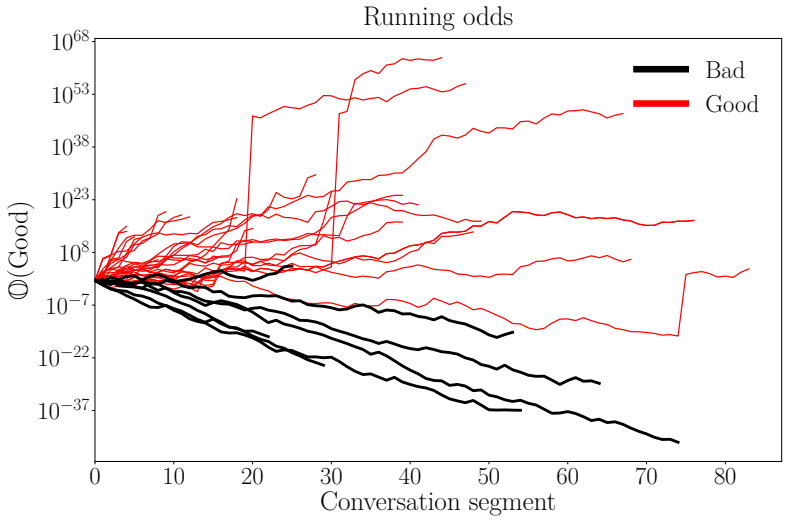


Figure 13: The running odds on the validation set.

The reader might be concerned by the size of the jumps in the updating scheme as seen on the validation set from Figure 13. This is because this scheme was trained on a wider variety of features, among which there would occasionally be outliers. Given the Gaussian kernel density estimates, this means model probabilities may be exceptionally close to zero, and hence division by such numbers in the updating formula may occur (sometimes under-flowing numerically to zero). The weight given to such outliers should

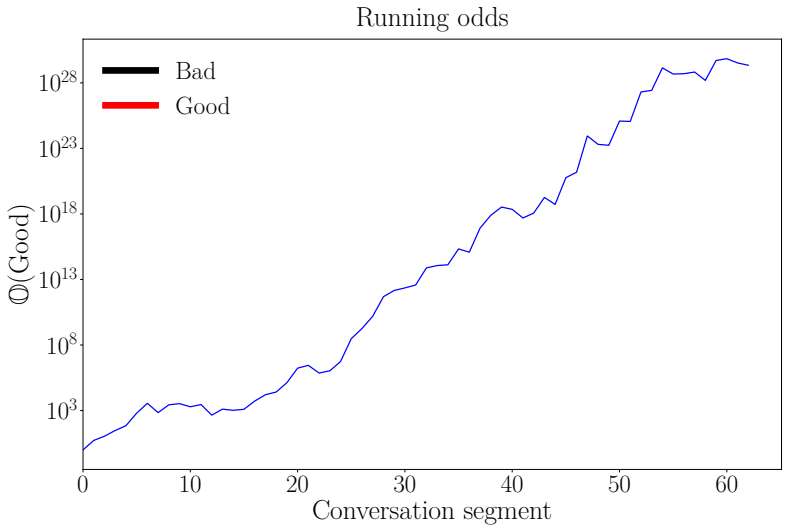


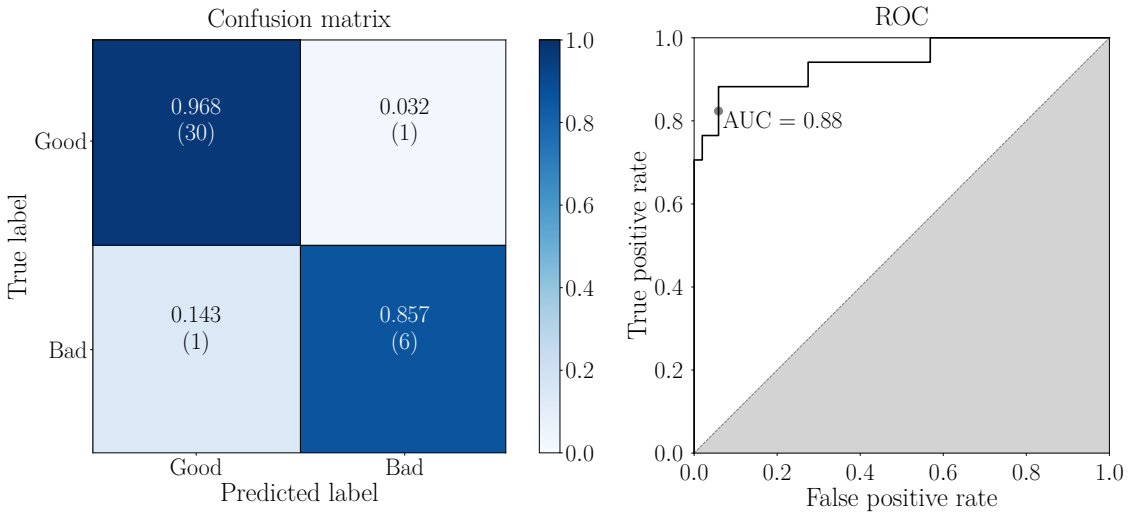
Figure 14: The running odds on a real clinical conversation which was classified correctly as a good conversation.

not be as large, and it is therefore best at this stage to cap the sizes of these influences.

4.4 AUC

The question that remains is how best to assess the performance of this classification scheme. It is important to note that if there is a significant imbalance in the data then global accuracy becomes a very inappropriate metric upon which to measure the performance. This is best illustrated by a counter example. Imagine a classifier that can classify 'bad' conversations with 70% accuracy, and a 'good' conversations also with 70% accuracy. The global accuracy for this classifier would be 70%. However, suppose only 10% of the incoming conversations were 'bad'. If this were the case then we could achieve a 90% global accuracy by the naïve classification that all conversations are always 'good'. Notice that with such a crude classification scheme, although we would have a higher global accuracy, we would be misclassifying all of the bad conversations, and hence defeating our original intention.

The important output from the running odds as in Figure 12 is not their actual values, but rather their ordering. In this scheme, after all the data were collected from a conversation, if the odds were greater than the initial prior, then the conversation was classified as 'good'. However, this corresponds to a 50% confidence that the conversation was 'bad' for it to be classified as 'bad'. Instead we might want the more stringent "only classify a conversation as 'bad' if the scheme is 99% confident", or similar. This will control the rate of true positives and false negatives. To illustrate the problem, it is useful to inspect the confusion matrix for the classification on the validation set, as shown in Figure 15a.



(a) The confusion matrix on the validation set.

(b) An example ROC curve achieved when only using 10% of the data for training.

Figure 15: The confusion matrix obtained for the validation set, and an example ROC curve obtained for a limited training set.

The value of the required confidence threshold for classifying a sample as a 'good' or a 'bad' conversation determines the ratio of the false positive rate to the true positive rate. Plotting these out against each other for differing confidence thresholds produces what is called a receiver operating characteristic (ROC) curve. We present an example ROC curve in Figure 15b, where the solid marker indicates the achieved performance, the diagonal line indicates the performance expected from random guessing, while the shaded region indicates a performance worse than random guessing. (The example ROC curve is only trained using 10% of the data to provide a more illustrative ROC curve and AUC value). A common alternative to the global accuracy for data sets with a class imbalance is the *area under the ROC curve*, commonly abbreviated as the AUC. In Figure 15b the AUC value is 0.88, although when using c60% of the data for training the AUC values were typically 0.95 or better.

5 Hidden Markov Model

5.1 Introduction

In this section, we are going to explain hidden Markov chain and its application to dialogue classification. Markov chain has "finite memory property", i.e. given the present, the future is independent of the past. Applications of Markov chain model range from queuing models, search engine and automatic text generation. Hidden Markov model assumes the system is determined by an underlying hidden Markov chain, but we can only observe consequences of the chain (O) , not the chain itself (Figure 16). An important application of Hidden Markov model is in modern speech recognition.

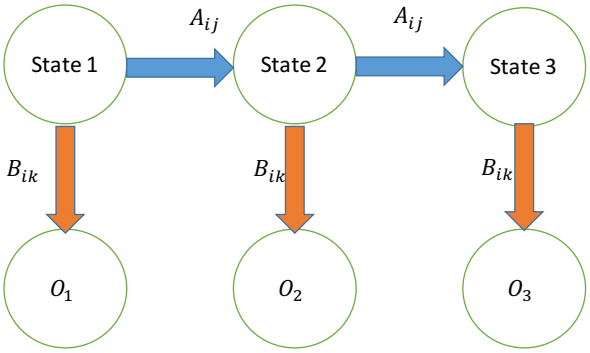


Figure 16: Diagrammatic illustration of the Hidden Markov model

The hidden Markov model we consider in this report is specified by the transition matrix A, the observation matrix B, and the initial distribution π. A, B and π are all row stochastic matrix (row sums to one) and they have the following interpretation:

- $A(i,j)$ is the probability that the next state is j given the current state is i,
- $B(i,k)$ is the probability of observing k given the current state is i,
- $\pi(i)$ is the probability that the initial state is i.

Starting from the initial distribution π, the chain develops according to the transition matrix A, while producing observation sequence O via the observation matrix B. Once we have fixed the number of hidden states (i.e. the size of the matrix A), it is possible to learn the parameters π, A, B from the observation sequence O by a maximal likelihood estimation. The algorithm is called Baum-Welch algorithm. We will use the implementation in the Python package *hmmlearn*.

In our problem of dialogue classification, we could think of each state represent the current state of the dialogue – it could be good, bad or neutral, and the observations are various features of one turn, e.g. average pause duration, number of words spoken etc. From this observation sequence, we could then learn the state sequence using

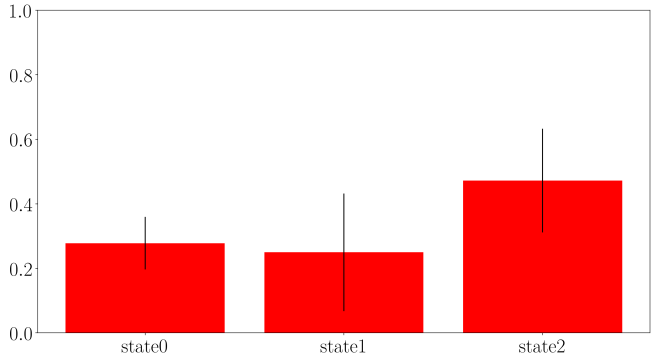


Figure 17: Average hidden state distribution for good conversations

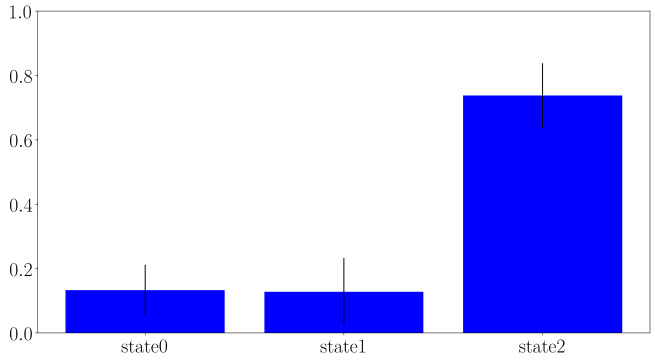


Figure 18: Average hidden state distribution for bad conversations

the package mentioned above. We expect the a good conversation will have a larger percentage of "good" stages, and a bad conversation will have a larger percentage of "bad" stages, so that good and bad training examples show different stage distributions. Given a new conversation consists of many turns, we could then predict the state of the dialogue, i.e. the hidden state in our hidden Markov model, and compare the state distribution to our training sets, this then gives us some indication of whether the testing dialogue is good or bad.

5.2 Result of state distribution of Good and Bad conversations

In Figure 17 and Figure 18, we show the distribution of hidden states for good and bad training examples. Comparing these two, we see that good conversations tend to have a larger proportion of hidden state 0 and 1, while bad conversations have much larger proportion of hidden state 2.

From this, given a new dialogue, we calculate its state sequence and hidden state distribution dynamically, then, the appearance of hidden state 2 may signal that the conver-

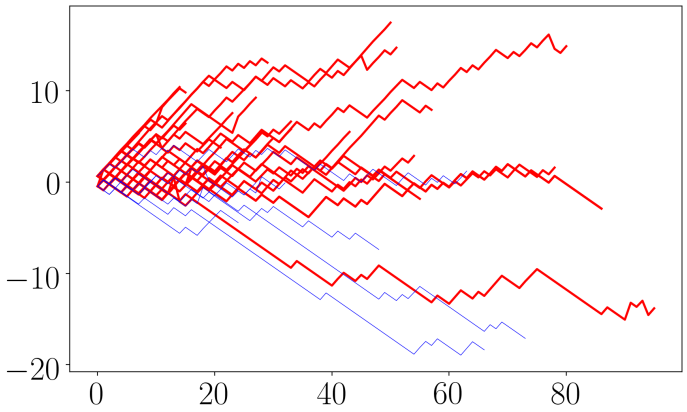


Figure 19: Running log odds plot for Bayesian updating scheme based on distribution of hidden states

sation is turning bad, while the absence of such is an indication of good conversation.

5.3 Using the hidden state distribution as a new feature and running odds classification

In the previous section, we saw the Bayesian updating scheme where running odds are calculated dynamically by using some features. We could think about the distribution of hidden states as new features, and try to implement a similar Bayesian updating scheme based on those features. The resulting running odds plot are shown in Figure 19.

We see that based features of hidden states distribution, the Bayesian updating scheme is able to classify correctly most of the good and bad conversations in the test set. Though the accuracy is not as good as Figure 12, the hidden Markov approach offers a potentially more explainable model that the hidden states could represent the "good", "bad" and "neutral" states of the turn.

6 Simple classifier based on formal statistical testing

6.1 Overview of statistical testing

Statistical testing is also known as hypothesis testing. It is typically formulated as testing of null hypothesis (H_0) against alternative hypothesis (H_1). The hypothesis could be any statement we want to test, for example suppose that we are interested in knowing whether certain parameter θ is statistically different from zero, then the test is simply $H_0 : \theta = 0 \quad vs \quad H_1 : \theta \neq 0$. or it could be testing whether two samples are from the same underlying distribution. In the context of the NHS dialogue classification problem, hypothesis testing can be used to test whether the sampling distribution of certain feature of the good dialogue is significantly different from the distribution of that feature of the bad dialogue.

In more formal sense, we define a *statistic* to be any function of the data sample $\{X_i\}_{i=1}^n$ drawn from a possibly unknown distribution. Let the statistic be $S = S(X_1, ...X_n)$, its distribution is called the *sampling distribution*. The basic idea of any hypothesis testing is to find a statistic such that, under the null hypothesis, the sampling distribution has a known form; that is, we can compute any probabilities of S. Then, by comparing the observed value of S with the probability distribution of S under the null hypothesis, we will be able to know how likely it is to observe S under the null hypothesis, if the observed S falls in the tail of its sampling distribution, which is unlikely, we will reject the null hypothesis, otherwise we say there is not enough evidence to reject the null hypothesis. In common statistical practice the line where "significance" is defined is typically the 95% quantile of the sampling distribution: $P(S \leq S_o) = 1 - \alpha$, with $\alpha = 0.05$. We can find S_α by inverting the distribution function of S.

Hypothesis testing has been built into many statistical software, what they typically report is the *p-value*, defined as $P(S > S_o)$, the probability that S is greater than S_o, the observed value of S. Then we can directly compare the p-value with the significant level of our choice, which may not be 0.05 depending on how tolerant we are to the types of error we might make. Using 0.05 as an example, if the p-value is below it then the test is significant we reject the null hypothesis, otherwise we do not.

6.2 Classifier based on K-S test

With the factors important to NHS (outlined in Section 2.1) in mind, one can design a simple classifier that uses the result of certain statistical test as a proxy for classification. In particular, the Kolmogorov-Smirnov test (K-S test) can be used to compare two samples, one from any dialogue and the other the reference dialogues with known label, and test whether they come from the same distribution. Here we use the good dialogues as reference because 1) they are more abundant than bad ones in the provided dataset

and so are more reliable, and 2) it is more natural to assume that any new conversation is good, that the patient is willing to listen, and reject only when there is sufficient evidence against. The idea of K-S test based classifier is that by recording feature values of an ongoing dialogue, and performing the test to test equality of feature distribution coming from the ongoing dialogue and that from the reference good dialogues, we get a p-value indicating whether the two distributions are significantly different — if the p-value is less than 0.05, then the ongoing dialogue is classified as bad, otherwise it is classified as good.

This method can deal with real-time learning naturally. One simply record feature values of the ongoing dialogue as it progresses in time, and perform the test at every time interval. However, the test is only reliable when there is enough data, i.e. when more feature values are collected, so a cautionary remark is that the p-values coming from the early stage of a conversation may not be indicative, even if some of them are below the 0.05 threshold, and one probably should wait longer before making any decisions. I will outline below how we can alleviate this problem.

Choosing which feature to use is critical to the success of this method, because an inappropriately chosen feature may at best not have the necessary power to differentiate between good and bad dialogues, and at worst be misleading. Generally speaking, features that have been identified as important in standard machine learning algorithms, such as lasso regression and random forest, are possible candidates of features in this method. Here I use two features, one is the ratio of the gap between speakers to the duration of the sentence immediately after the gap (g/d), and the other is the natural logarithm of the number of words in sentences.

Back to the problem of unstable p-values, one idea is to use multiple features and perform the K-S test simultaneously. This way the features act as a committee and whether a conversation is good or bad is determined by all members of the committee, so a call handler will be more confident if most members report consistent p-values, in the sense that they are all above or below 0.05, and less so otherwise. It is also easy to parallelize since all tests are carried out independently. Another idea that might mitigate this issue, from the perspective of an individual test, is to put more weights on the later part of the dialogue by repeating (with some noise added) the last few feature values collected so far before performing the K-S test. Intuitively, as the dialogue is ongoing, the later feature values should have stronger signal of how the dialogue is going to develop next, than those collected earlier, and hence they should be weighted more. For example, suppose we have collected 10 feature values from an ongoing dialogue, and we wish to perform the test at this point, we can replicate the last 5 values with added noise to break ties, and use the total 15 feature values from both real and artificial data to perform the test, rather than just 10 from the real data.

This second idea provides opportunities for research into more efficient ways of utilizing the data. For example, a fancier way to augment the dataset than repeating the last few feature values could be constructing the empirical cumulative distribution function from

the last part of feature values and draw random variates from it.

6.3 Results

There are 28 bad and 88 good dialogues in the training set. One of the bad dialogues is a monologue, so it is not used. 80% of the good dialogues were used to construct reference feature distribution, so the total number of dialogues to classify is 45 (27 bad, 18 good). Figure 20–22 show the running p-values of the K-S test for the 27 bad dialogues, for two features g/d the gap-duration ratio and `lwords`, the logarithm of words in sentences. If p-value is below 0.05, then the testing distribution is believed to be different from the reference distribution, and so the prediction is bad; otherwise the prediction is good. As the conversation proceeds, we expect the running p-values fall below 0.05, and we see that for many this is the case, such as B2, B3, B4, B8[5] and so on. However, we also see that for some dialogues one feature is better than the other, for example, in B20, B22, B24 g/d is better than `lwords`, and in B12, B13 `lwords` is better than g/d; there are some dialogues for which neigher of the features are good, for example B7, B10, B11. For B7 it seems confusing because the p-values for `lwords` stayed below 0.05 for some 20 gaps, and then went up again. For B10 and B11 it is not surprising because there are not many gaps and the test would not be stable with only a few data points.

Figure 23–24 show the running p-values of the K-S test for the 18 good dialogues. In this case, we expect the running p-values to be above 0.05, because then there is not sufficient evidence to believe the dialogue is bad. Here one can see that g/d is better than `lwords` in most dialogues. In G5 and G51, even though the p-values from both features indicate that the dialogue is good in the end, the `lwords` p-values stayed below 0.05 for quite some time near the end and by which point the call handler would probably have made the wrong decision. One exception is G14, in which case `lwords` is better than g/d. G54 is particularly interesting, because the two features give conflicting results, with g/d suggesting good dialogue whereas `lwrods` suggesting bad. Perhaps G54 contains characteristics of both good and bad dialogue, and it would be interesting to have a group of people consider this dialogue and see what proportion think it is good.

Table 1 shows the classification results based on all feature values collected for each testing dialogue, and for each feature. From this we can calculated various measures of the classification, some of which are listed in Table 2. We can see that g/d is a better than `lwords`, which confirms the observations in the figures above. Accuracy measures the overall misclassification rate on the whole dataset. Precision is the proportion of actual bad in all that are classified as bad; that is, out of all conversations classified as bad, how many are actually bad? Recall (or true positive rate), is the proportion of actual bad over all bad conversations in the data, so a high recall means the probability of detecting bad conversations is high.

[5] I use B2 to denote the second bad dialogue, G1 for first good dialogue and so on.

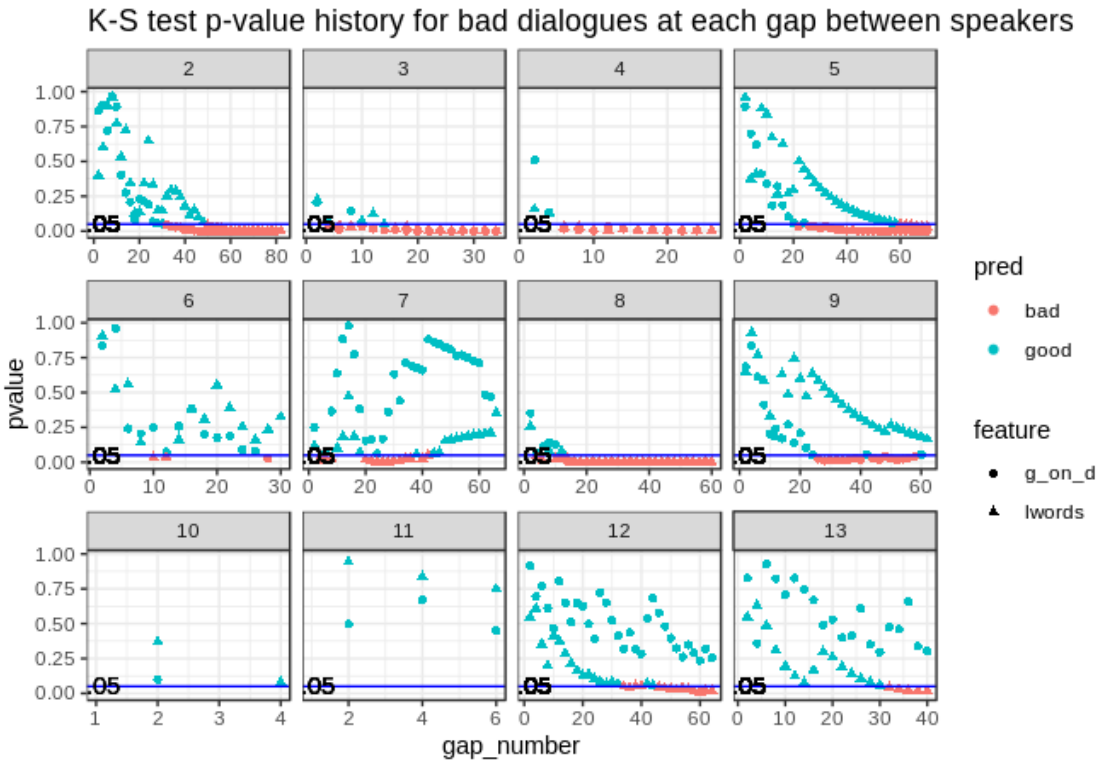


Figure 20: Running p-values as a function of gaps, for both features g/d (circle), the gap-duration ratio and `lwords` (triangle), the log of number of words in sentences. Predicted labels are colored differently depending on if they fall below the horizontal 0.05 reference line (blue). Figures 21, 22, 23, 24 show the same for the remaining 15 bad dialogues and all 18 good dialogues.

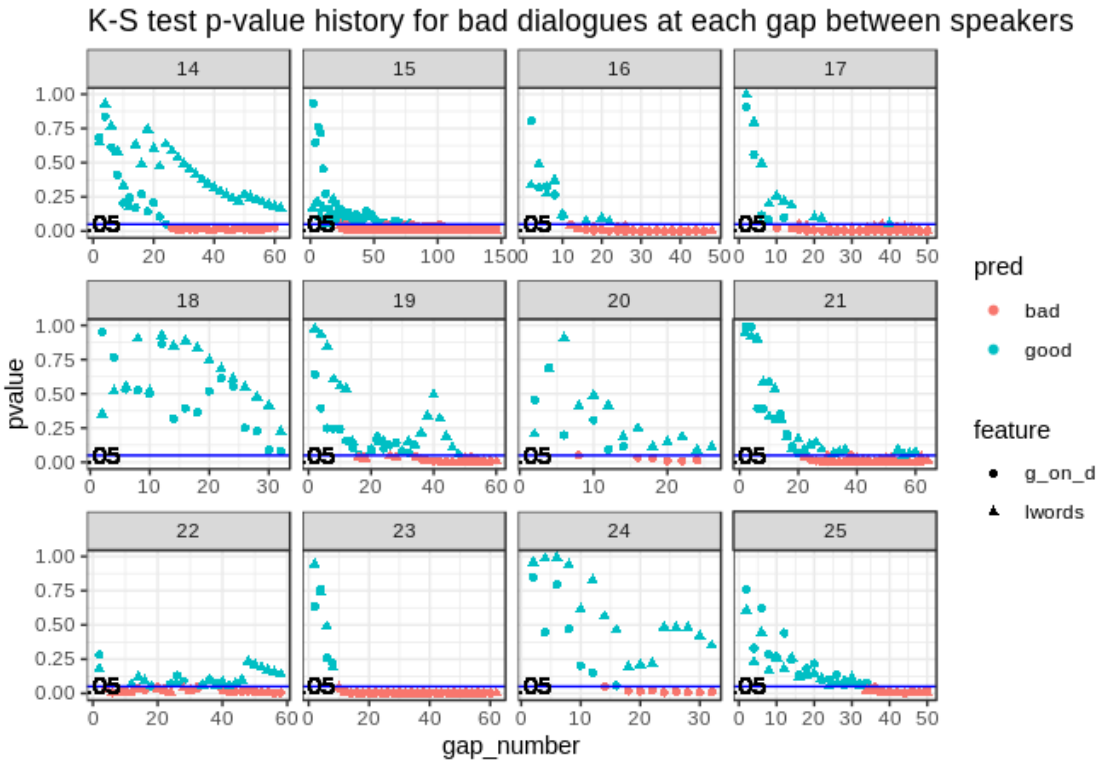


Figure 21: Part 2 of bad dialogues.

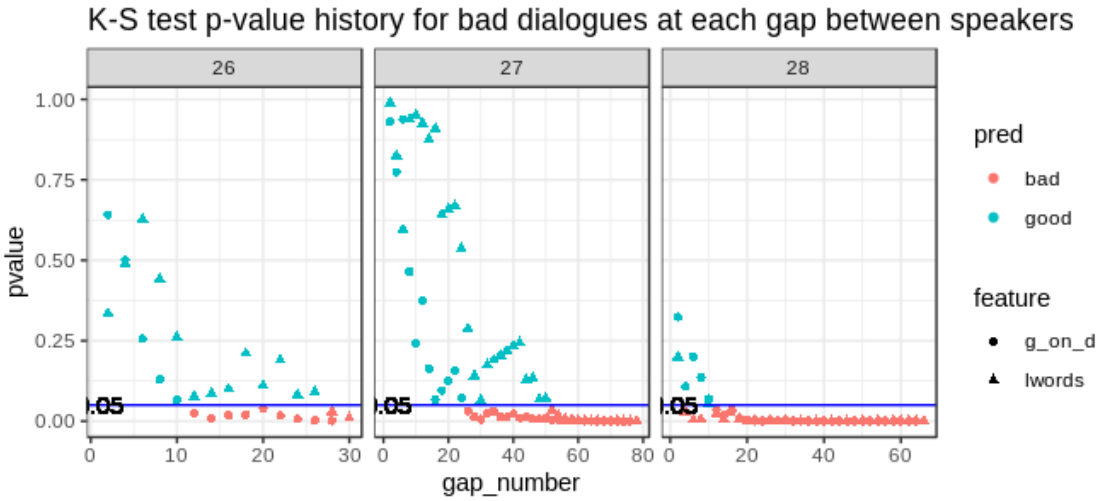


Figure 22: Part 3 of bad dialogues.

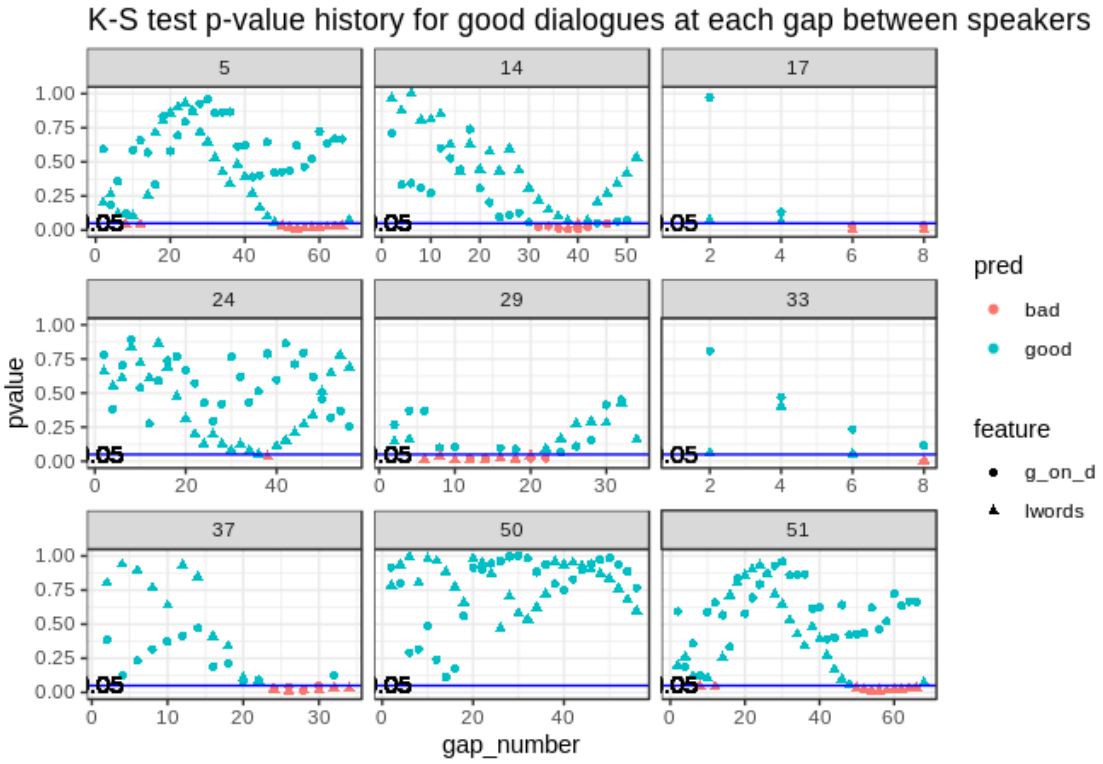


Figure 23: Part 1 of good dialogues.

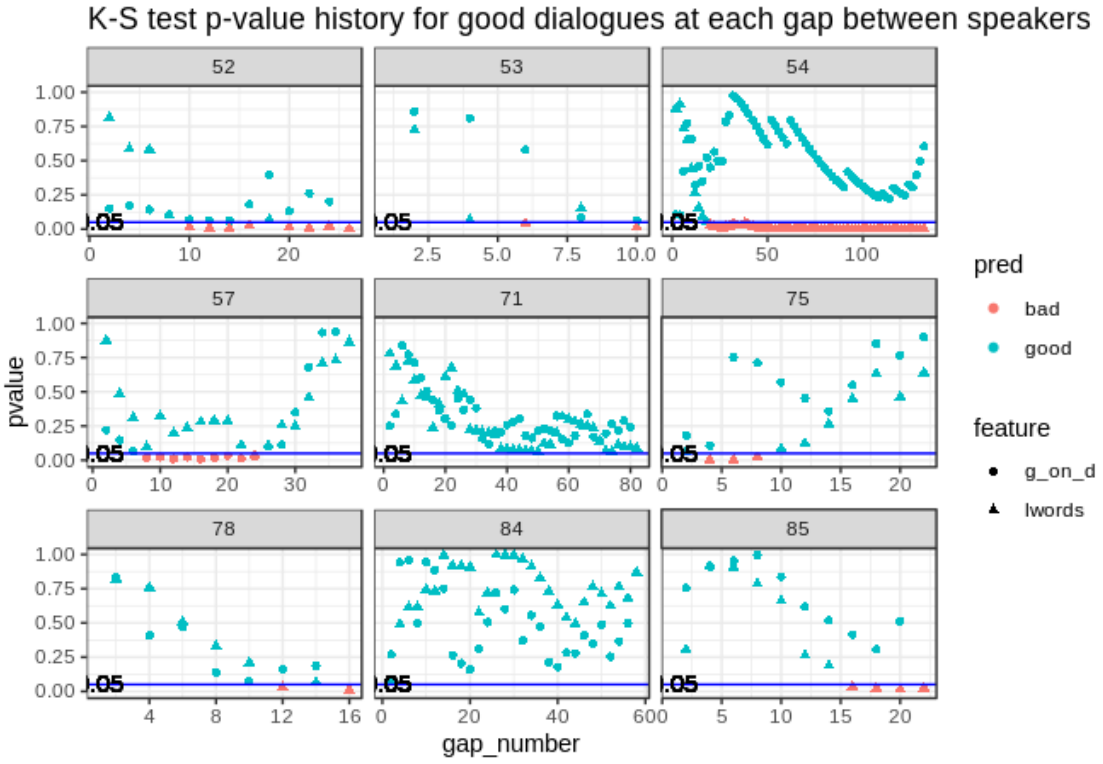


Figure 24: Part 2 of good dialogues.

Here we also tried the idea mentioned in section 6.2, on improving stability of the test when there is insufficient data. For each test carried out during the course of the conversation, the last feature value collected thus far was duplicated with some random noise, the classification result with this modification is shown in the last two rows of Table 2; and one can see that all performance measures are improved for both features; in particular using g/d, the augmented version with the last feature value duplicated has improved the classification a lot compared with using just the plain g/d values from real data.

		actual		Feature
		G	B	g/d
pred	G	17	7	
	B	1	20	

		actual		Feature
		G	B	`lwords`
pred	G	10	10	
	B	8	17	

Table 1: Confusion matrix of the classification, based on all feature values till the last gap of each conversation.

Feature	accuracy	precision	recall	F1 score	DuplicateLast
g/d	0.82	0.95	0.74	0.83	N
`lwords`	0.60	0.68	0.63	0.65	N
g/d	0.89	1.0	0.81	0.90	Y
`lwords`	0.62	0.69	0.67	0.68	Y

Table 2: Performance measure of the classification results. The first two rows correspond to the results in Table 1. The last two rows show the result of duplicating the last feature value of the testing dialogue before performing the K-S test. The F1 score is the harmonic mean of precision and recall.

6.4 Conclusion

In this section we gave an overview of hypothesis testing and how it can be done in a statistically formal way. I showed that K-S test can be used as a proxy for classifying good or bad dialogues in real time, by comparing feature distribution of gap-duration ratio and words in sentences with the corresponding feature distribution obtained from the reference dialogues. As with many machine learning methods, choosing good features is important to the success of this method and the gap-duration ratio is shown to perform better than words in sentences. While this test-based method may not be stable in the early stage of the conversation due to lack of sufficient data, I showed that by augmenting the feature values to include artificial data, such as duplicating the last feature value with noise, can greatly improve the classification result.

7 The Scorecard Method

7.1 Data Description

Data for 115 dialogues were provided. Subject matter specialists classified 88 of them as being good dialogues based on content information. The remaining 27 (one monologue excluded) were classified as bad dialogues. The dialogues were first electronically recorded, and then transcribed into digital text form using the IBM Speech to Text service. The resulting data set has the following structure: The left-most column contains the transcribed single words, which were masked using ordered numbers by the data provider; each dialogue data is expressed as a quartet: Column 1 and 2 contain the start and end time of a spoken word, respectively; Column 3 contains a class label for the speaker; Column 4 contains the time which elapsed (gap length) when one of the speakers stops speaking and another speaker starts to speak.

7.2 The Scorecard Approach

Here, we investigated the usefulness of the scorecard method [2] for classification of dialogues. In this method, we first construct partitions in an informative feature space using a classification tree approach. These partitions produce regions that could be defined as "good" or "bad", depending on a majority vote. After this, scores based on the logarithm of odds for a good dialogue can be assigned to these regions - positive score for a "good" region, and negative score for a "bad" one. By characterising the trajectory of a dialogue as a sequence of coordinates (at 30-second intervals) in the feature space, a dialogue can be assigned a cumulative score. The cumulative score at any one of the time intervals before dialogue completion can be assessed, allowing comparison with the cumulative score of the completed dialogue. If the scorecard method is effective, the cumulative score distributions for the good and the bad dialogues should have little overlap, and an intuitive threshold such as 0 should separate the good dialogues (mostly positive cumulative score) from the bad ones (mostly negative cumulative score).

7.3 Feature Extraction

We extracted three features from the raw data. The first one is the speaking rate, which is the total number of words spoken by all speakers (N) divided by the duration of the dialogue (T). For each dialogue, we can find N as the number of rows in an $N \times 4$ data matrix, and T is computed as the difference between the end time of the last spoken word and the start time of the first spoken word. The unit of measurement is words per second.

The second feature is the turn-taking rate, which is the number of turns (i.e. one speaker

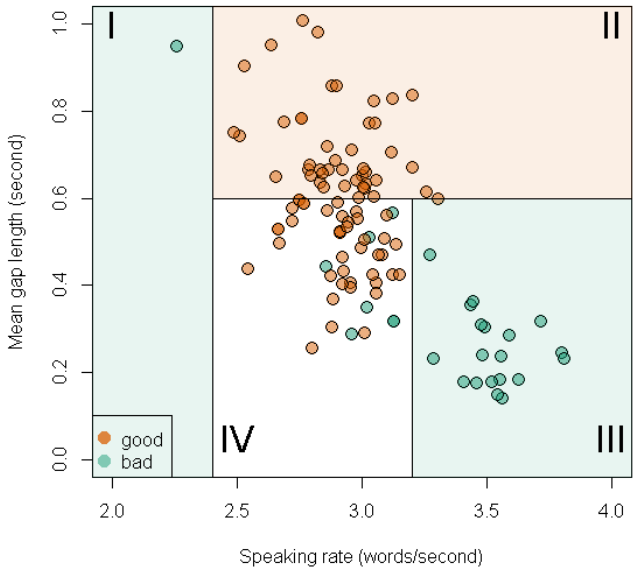


Figure 25: Scatter plot of completed dialogues in the mean gap length $\times$ speaking rate feature space. The orange (II) and green (I,III) regions indicate the space where predominantly good (orange) and bad dialogues (green) are found, respectively. Region IV consists of mixtures of good and bad dialogues.

stops speaking and another begins; U) divided by the duration of the dialogue (T), multiplied by 60, so as to have minute instead of second as the time unit. The unit of measurement is turns per minute.

The third feature is mean gap length, which is a weighted average of the gap length (G) between every turn U in the dialogue:

$$\text{Mean gap length} = \sum_{i=1}^{U} w_i G_i,$$

where w_i is the number of the words spoken by a speaker the the ith turn, divided by the total number of words spoken, N. The unit of measurement is seconds.

The distribution of the dialogues in feature space suggests good and bad dialogues can be associated with particular regions in the feature space. In Figure 25, we see that bad dialogues can be associated with two regions: (i) speaking rate ≤ 2.4; (ii) speaking rate ≥ 3.2 and mean gap length ≤ 0.6. Good dialogues are associated with the region speaking rate > 2.4 and mean gap length > 0.6. For dialogues falling in the region defined by $2.4 <$ speaking rate < 3.2 and mean gap length ≤ 0.6, a region at turn-taking rate ≤ 4 is associated with good dialogues (Fig. 26); a final partition using speaking rate generates regions associated with bad dialogues (speaking rate ≥ 3), and good dialogues (speaking rate < 3). For simplicity, the cut-offs for the partitions were heuristically determined. They can be inferred more formally using algorithms for classification and regression

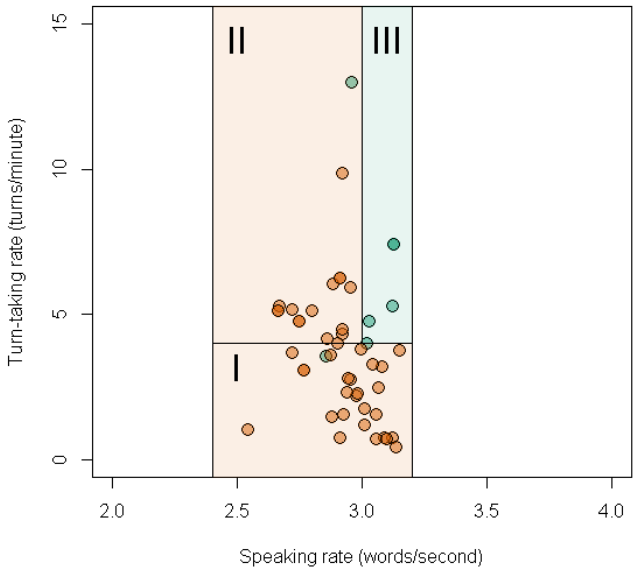


Figure 26: Scatter plot of completed dialogues (from Region IV of Fig.25) in the turn-taking rate $\times$ speaking rate feature space. The orange (I, II) and green (III) regions indicate the space where predominantly good (orange) and bad dialogues (green) are found, respectively.

tree such as C5.0 [3] and GUIDE [4].

7.4 Dialogue Trajectories and their Scores

Having established the usefulness of the three features, we next explored the trajectories of the dialogues at time intervals of 30 seconds.

As a dialogue traverses through the feature space for every 30-second time segment, a score can be assigned depending on which region of the feature space it is located (Fig. 27). The score assigned to a region can be based on the logarithm (base 10) of the odds for a good dialogue:

$$\text{Score} = \log_{10}\left(\frac{N_g + c}{N_b + c}\right),$$

where N_g and N_b are the number of (completed) good and bad dialogues in a region, respectively. The constant c is a pseudovalue added to address the situation where N_g or N_b is zero, or very small. For regions defining good dialogues, $c = 1$; for regions defining bad dialogues, $c = \sqrt{N_g/N_b}$ to adjust for the unbalanced class ratio in the present data set where there are 3 times fewer bad dialogues compared to good dialogues.

Thus, dialogues which spend a lot of time in "good regions" are expected to have large, positive cumulative scores; while those that spend more time in "bad regions" are ex-

	Dialogue class	
Cum.score	Bad	Good
-ve	24	5
+ve	3	83

Table 3: 2×2 contingency table for dialogue class and cumulative score sign. The sign-based cumulative scorecard has sensitivity of about 89% and specificity of about 94%.

pected to have large, negative cumulative scores. Classification of a dialogue can then be based simply on the sign of the cumulative score.

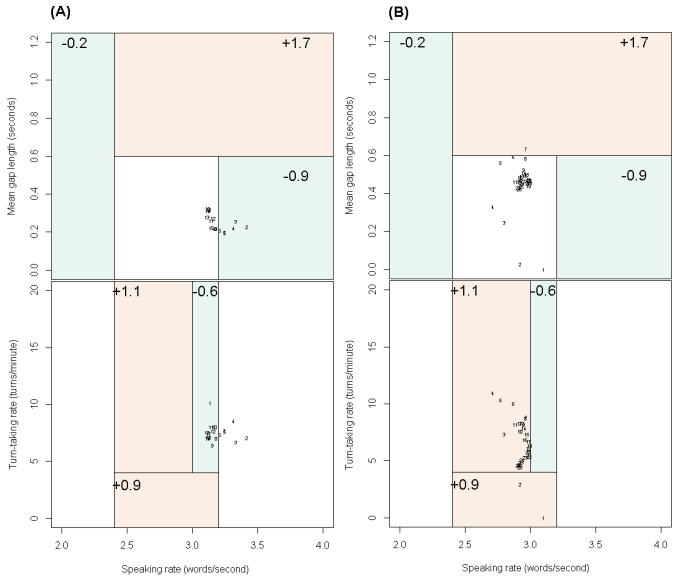


Figure 27: Trajectories of a bad (A) and a good (B) dialogue in the feature space. Note that the trajectory of the bad dialogue lingered in the "bad" regions (green), whereas that of the good dialogue moved through mostly "good" regions (orange). The numbers index the first,second,etc. 30-second segment of a dialogue. Scores are assigned to specific regions in the feature space, and a cumulative score is associated with each dialogue based on the trajectory history.

The dialogue trajectories with their cumulative scores can be visualised to study their behaviour over time (Fig. 28). In general, for the completed dialogues, good dialogues are characterised by positive cumulative scores, and bad ones by negative cumulative score (Fig. 29). Thus, a simple sign-based scorecard attains sensitivity (focus on detecting bad dialogues) of 89% and specificity of 94% (Table 3).

Table 4 shows the joint distribution of the sign of the final cumulative score, and the cumulative scores after 2, 5 and 10 minutes of a dialogue. Note that all bad dialogues which have negative final cumulative score already have negative cumulative score after 2 minutes. Most of the good dialogues which have positive final cumulative score have a positive cumulative score after 5 minutes. Figure 30 shows the distribution of

Time segment(min)	Score	Final cumulative score			
		Bad		Good	
		-ve	+ve	-ve	+ve
2	-ve	24	0	3	13
	+ve	0	3	2	70
5	-ve	23	0	5	5
	+ve	1	3	0	78
10	-ve	24	0	5	4
	+ve	0	3	0	79

Table 4: Agreement of sign of cumulative score at three time segments (2,5,10 minutes) with sign of final cumulative score for bad and good dialogues.

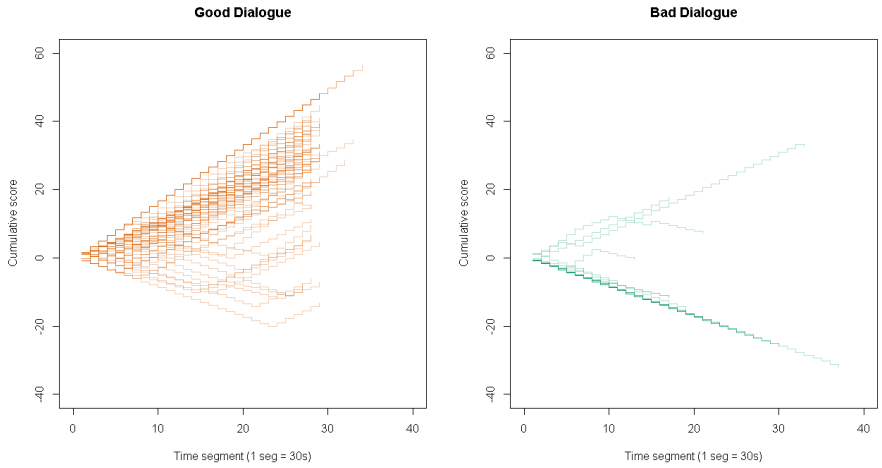


Figure 28: Score trajectories for good (orange) and bad (green) dialogues.

the dialogues in feature space after 2 and 5 minutes elapsed.

7.5 Discussion

In a dialogue, participants alternate between the role of a speaker and a listener, known as turn-taking, with variable duration of pauses. Interestingly, turn-taking as an informative feature of dialogues is anticipated by Sacks et al. [5], who suggested that conversations can be characterised by turn-taking patterns without the need for context:

> "We have found reasons to take seriously the possibility that a characterization of turn-taking organization for conversation could be developed which would have the important twin features of being context-free and capable of extraordinary context-sensitivity (p.699)".

The partition cut-offs in the feature space appear to have useful interpretation. Firstly,

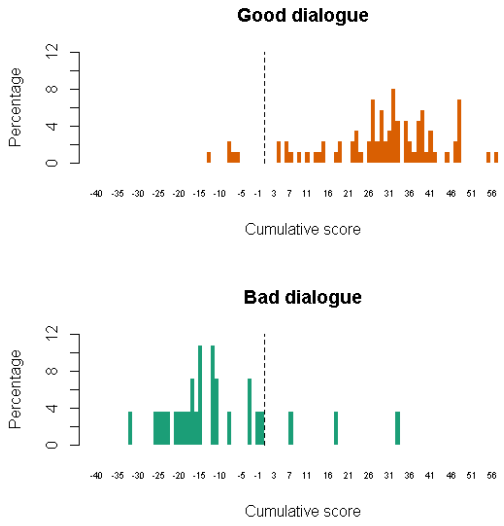


Figure 29: Distribution of cumulative scores for good and bad conversations. About 6% of the good dialogues (5/88) had negative final cumulative score. For bad dialogues, about 11% (3/27) had positive final cumulative scores. Colour annotation: orange = good dialogue; green = bad dialogue.

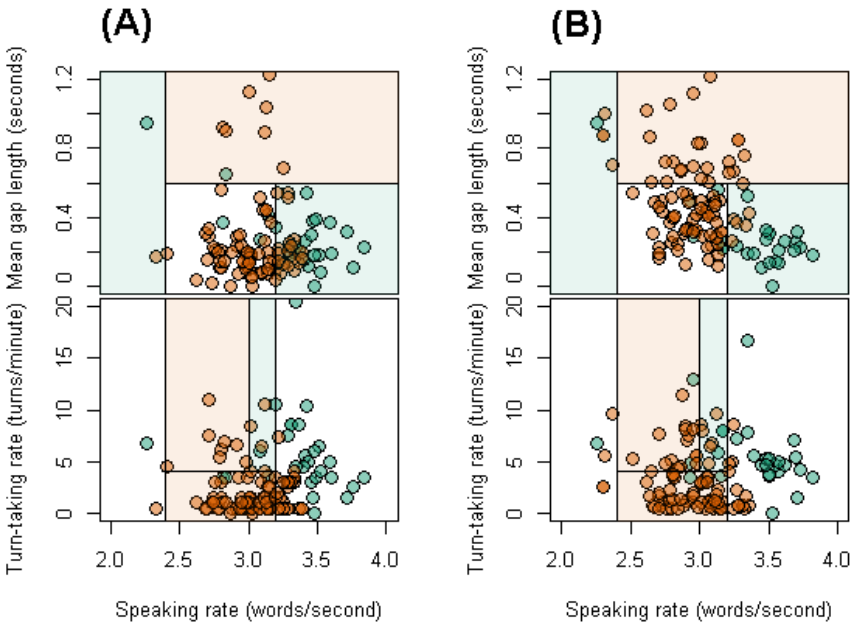


Figure 30: Distribution of dialogues in feature space after (A) 2 minutes; (B) 5 minutes. Colour annotation: orange = good dialogue; green = bad dialogue.

the "bad" region defined by speaking rate of 3.2 or more, with mean gap length below 0.6 (Region III in Fig. 25) characterises dialogues where speakers spoke rapidly with less pause between turns. This seems to be reasonable, as in bad dialogues, speakers tend to speak rapidly, and do not pause long enough to reflect on the spoken content. On the other hand, the region defined by mean gap length above 0.6 and speaking rate between 2.4 and 3.2 is "good" (Region II in Fig. 25) - the speakers were unhurried, and not reticent, which are characteristics of good dialogues. The "bad" region defined by speaking rate below 2.4 (Region I in Fig. 25) contains a single bad dialogue (ID 11), which is the shortest one (53 seconds). Presently, we feel that creating this partition to accomodate a single bad dialogue is reasonable, since this region characterises dialogues where speakers are hesitant. Getting more training samples with shorter dialogue duration will better inform the need for this partition.

For modest-paced dialogues (speaking rate between 2.4 and 3.2, mean gap length below 0.6), consideration of the turn-taking rate turned out to be informative. Modest-paced dialogues which had lower turn-taking rate (4 per minute) were mostly good ones, which is reasonable, as lower turn-taking rate is likely associated with better comprehension or willingness of a speaker or speakers to engage in a meaningful dialogue. For dialogues with turn-taking rate higher than 4 per minute, dialogues spoken with more rapidity (speaking rate of 3 or more) were mostly bad conversations.

For bad dialogues, the finding that the sign of cumulative score at 2 minutes agreed with that of the final cumulative score suggests the possibility of early intervention, such as introducing an experienced moderator into the dialogue to help de-escalate a difficult situation.

It may be possible to enrich the data set with additional data such as the time of the dialogue (morning or evening) and the gender of the initiator of the dialogue (based on vocal properties). This could enable a segmented scorecard (creating different scorecards for different genders or time of dialogue) approach which can further improve classification accuracy.

Finally, we have eschewed a more formal approach of evaluating the performance of the proposed scorecard method (i.e. dividing the data set into training and test sets, performing cross-validation and examining ROC curves), because the small sample size, particularly for bad dialogues, limits the generalisability of any trained classifier. Furthermore, other than the state of mind of the speakers, the speaking rate is known to be influenced by multiple factors such as age, sex, dialect region, native speaker, familiarity with other speaker, and topic of conversation [6], all of which are unknown in the present data set. It is necessary for future training sets to include samples that adequately cover variations in these demographic variables to avoid over-fitting problems in a trained classifier.

7.6 Summary

Electronically transcribed, context-free dialogues have three standard features that are useful for the discrimination of good and bad dialogues. These features are the speaking rate, the turn-taking rate, and mean gap length. In this work, we first partitioned the feature space into "good" and "bad" regions, using the distribution pattern of completed dialogues in the feature space. Subsequently, we assigned a log of odds score in favour of good dialogues to these regions, with positive scores for "good" regions and negative scores for "bad" regions. We then developed a scorecard assessment, where each dialogue was given a cumulative score based on its trajectory in the feature space over 30-second time segments. We found that, as early as two minutes after a dialogue had begun, the cumulative scores of bad dialogues were mostly negative. This finding opens the interesting possibility of early intervention to influence the outcome of a dialogue.

8 Comparison of Methods

Here we directly compare three of the classifiers presented above:

- the Bayesian scheme with odds derived from the histograms of gap length and duration (see Section 4),
- the Bayesian scheme using the proportion of time spent in the different Markov states, obtained from the same features (see Section 5), and
- the classifier based on the p-value of a Kolmogorov-Smirnoff test, which estimates the probability of the test sample belonging to the class of 'good' dialogues, by considering the ratio of the gap between speakers to the duration of the sentence immediately after the gap (g/d), and the natural logarithm of the number of words in sentences (see Section 6).

31 'good' and 8 'bad' conversations were reserved for validation, out of 116 conversations in the full dataset. Of the remaining conversations 60% were used for training of the model, whereas 40% were used for testing. To see how the model performs on unseen data, we assessed its performance on the validation set. For this set we plot the evolution of the value of 1 – AUC in Figure 31, evaluated as the number of conversation segments seen increases over time; in other words, at each segment we compute the AUC assuming that we have only seen all the previous conversation segments, but not the subsequent ones. This not only allows us to compare different classification methods, but also enables us to evaluate the number of conversation segments each method needs to be able to classify the dialogue with sufficient accuracy.

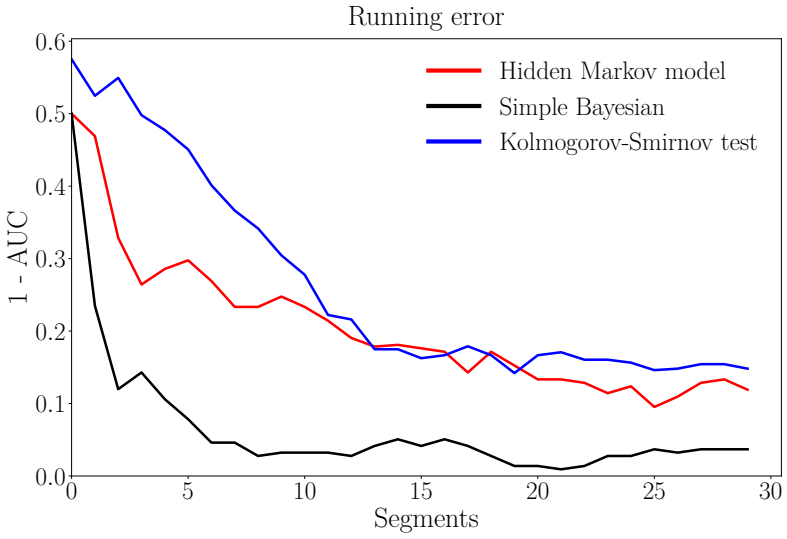


Figure 31: Performance of three different classifiers for the validation set against the number of conversation segments seen

The Bayesian method using gap distribution achieves an AUC of over 0.9 within five conversation segments. This is dramatically quicker than the other two approaches, which even at 30 segments do not attain comparable performances. Our results suggest that a simple Bayesian odds updating scheme informed by explicit characteristics of the data can quickly and effectively differentiate between the two conversation types.

9 Summary, Conclusions and Future work

We used a variety of approaches to address this problem. Data exploration revealed significant differences between the two categories of conversation for which data was provided. We then attempted to classify the conversations as 'good' or 'bad', using a number of approaches trained on a subset of the data and tested on a separate subset: Bayesian updating scheme; a classier based on the K-S statistical test; a hidden Markov model; and a scorecard method. All methods were able to generally distinguish between the 'good' and 'bad' conversations. However, caution must be taken in using these methods on actually NHS call data where the differences between calls may be far more subtle.

The next steps would be to apply similar exploratory data analysis to actual calls to NHS that have been categorised as good or bad. This would enable determination of the most suitable metrics with which to classify the calls. Then, the methods described in this report can be re-applied to the new data with the most suitable metrics, and the methods compared. There is great potential for this approach to lead to the development of a real-time classifier of an exchange.

10 List of Acronyms

References

[1] Peter Grindrod. *Mathematical underpinnings of analytics: theory and applications*. OUP Oxford, 2014.

[2] D.J. Hand. Good practice in retail credit scorecard assessment. *Journal of the Operational Research Society*, 56:1109–1117, 2005.

[3] J.R. Quinlan. *C4.5: Programs for Machine Learning*. San Mateo, California: Morgan Kaufmann Publishers, 1993. ISBN 1558602380.

[4] W.Y. Loh. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12:361–386, 2002.

[5] H. Sacks, E.A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735, 1974.

[6] J. Yuan, M. Liberman, and C. Cieri. Towards an integrated understanding of speaking rate in conversation. *The Ninth International Conference on Spoken Language Processing (INTERSPEECH-2006)*, pages 541–544, 2006.