Published in final edited form as:
*Proc IEEE Workshop Autom Speech Recognit Underst*. 2008 April 3; 1: 378–384. doi:10.1109/ICCV.2003.1238369

# Supervised and Unsupervised Feature Selection for Inferring Social Nature of Telephone Conversations from Their Content

**Anthony Stark**,
Biomedical Engineering, OHSU, Portland, USA starkan@ohsu.edu

**Izhak Shafran**, and
Biomedical Engineering, OHSU, Portland, USA shafrani@ohsu.edu

**Jeffrey Kaye**
Biomedical Engineering, OHSU, Portland, USA kaye@ohsu.edu

## Abstract

The ability to reliably infer the nature of telephone conversations opens up a variety of applications, ranging from designing context-sensitive user interfaces on smartphones, to providing new tools for social psychologists and social scientists to study and understand social life of different subpopulations within different contexts. Using a unique corpus of everyday telephone conversations collected from eight residences over the duration of a year, we investigate the utility of popular features, extracted solely from the content, in classifying business-oriented calls from others. Through feature selection experiments, we find that the discrimination can be performed robustly for a majority of the calls using a small set of features. Remarkably, features learned from unsupervised methods, specifically latent Dirichlet allocation, perform almost as well as with as those from supervised methods. The unsupervised clusters learned in this task shows promise of finer grain inference of social nature of telephone conversations.

## I. Introduction and Motivation

With the growing popularity of voice interface in portable devices, there is an increasing demand for context-sensitive computing and context-sensitive user interfaces. For example, during a business call, a user may wish to access his calendar or his office emails. Likewise, during a personal call, he may wish to access his twitter messages or his facebook page. Such context-sensitive interfaces can help reduce cognitive load under challenging circumstances (e.g., driving) as well as for older adults. Apart from facilitating efficient interaction with others, the ability to automatically infer the social nature of conversations can provide new tools for social psychologists and social scientists to study and understand social life of different subpopulations under different situations.

The goal of this work is to characterize social engagement in older adults so that we can better understand its ameliorating effect on cognitive decline [1]. While social engagement is multifaceted, older adults, who are often less mobile, rely on telephone conversations to maintain their social relationships. A recent survey by Pew Research Center found that among adults 65 years and older about nine-in-ten talk with family or friends every day and more than 95% of adults use landlines telephones for all or most of their calls [2].

Given the above premise, we focus our attention on studying social interactions over landline telephones among older adults. To facilitate such a study, we collected telephone conversations from a small group of older adults for several months, after obtaining their consent, as described in Section II. Note, this corpus is unlike publicly available corpora

such as Switchboard and Fisher in that the conversations are natural and cover a wider range of topics from everyday life with more conversational partners.

Teasing apart the nature of spoken discourse is an active area of research with topics such as affect recognition [3], recognition of speaker characteristics [4], determination of dialog acts [5], [6], [7], content-driven topic modeling [8], [9], [10] and the role of small talk versus substantive conversation in social well-being [11]. Of particular interest to our own work are the studies conducted over the Switchboard telephony corpus [12]. This corpus contains telephone recordings between pairs of individuals, where participants were asked to draw from a particular theme of social conversation (e.g. music, crime, air pollution) when chatting over the telephone. The presence of well-defined labels has led to several Switchboard topic classification studies [13], [14]. Despite imperfect automated speech recognition, topics were discriminated with a surprisingly good accuracy. While Switchboard has been undeniably useful in probing the topic classification problem, it is limited by the fact that all conversations consist of social discourse with an unfamiliar contact.

Everyday telephone conversations, on the other hand, are highly diverse in nature, ranging from business-oriented calls to highly personal calls. As an initial step towards characterizing these calls, we investigate a coarse classification between business-oriented calls and calls from/to residential lines. The privacy constraints on the corpus preclude human transcriptions and annotation of the data. Instead, we rely on side information collected from online sources. This classification task provides a mechanism to learn features related to lexical choice and speaking style that signal formal and informal social contexts. Such features may be useful in characterizing the diverse calls observed in everyday life. For this purpose, as described in Section III, we investigate a variety of features, extracted solely from the content, content that is obtained from automatic transcription and is errorful. We examine unsupervised methods for feature selection to utilize the large number of calls without side information. In Section IV, we analyze the errors incurred by our classifier and show that the confidence of the classifier can be quantified.

## II. Corpus of Everyday Telephone Conversations

### A. Overview

Our corpus consists of 12,067 digitized landline telephone conversations. Recordings were taken from 10 volunteers over a period of approximately 12 months. Subjects were all native English speakers recruited from the greater Portland area in US. Audio captured from the USB recording devices was processed off-line with an ASR system. Apart from recorded speech, our corpus includes a rich set of meta data, including subject ID, time of day, call direction (incoming vs outgoing), duration and DTMF signaling tones, which we have not exploited in this study.

For this initial study, we discard conversations with less than 30 automatically transcribed words primarily to get rid of spurious and noisy recordings related to device failure and incorrectly dialed telephone numbers. Moreover, short conversations are less likely to provide enough social context to be useful. In addition to dropping short conversations, we also pre-process our transcripts with a short stop-word list. This list was created from the ten most numerous words in the corpus. Words appearing on this list were eliminated from the transcripts.

Of the 8,558 available conversations, 2,728 were identified as residential conversation and 1,095 were identified as business conversation. This left 4,395 unlabeled records, for which

the reverse-phone-lookup was either inconclusive or for which the phone number information was missing and/or improperly recorded. From the labeled records we created a balanced verification set containing 164,115 words over 328 conversations. The remainder was used to create a balanced training set consisting of 866,696 words over 1,862 conversations. For the balanced training set, we randomly selected 931 of the available 2564 residential training examples. Empirically, we found that dropping the 1,631 additional training records had negligible impact during cross validation experiments. The detailed breakdown of the corpus is give in Table I.

### B. Automatic speech recognition

Our ASR system is structured after the IBM's conversation telephony system which gave the top performance in 2004 evaluation of speech recognition technology by National Institute of Standards and Technology [15]. The acoustic models were trained on about 2000 hours of telephone speech from Switchboard and Fisher corpora [12]. The system has a vocabulary of 47K and uses a trigram language model with about 10M n-grams, estimated from a mix of transcripts and web-harvested data. Decoding is performed in three stages using speaker-independent models, vocal-tract normalized models and speaker-adapted models. The three sets of models are similar in complexity with 8000 clustered pentaphone states and 150K Gaussians with diagonal covariances. Our system does not include discriminative training and performs at about 24% on NIST RT Dev04. While our ASR system was trained on telephony speech, the training data differs from our corpus in diversity of content and speaking style as noted earlier. We cannot report a direct measure of transcription accuracy on our corpus since we lack human-generated reference transcripts.

## III. Classification experiments

### A. N-gram Features

To form an experimentation baseline, we examine the use of n-gram modeling. N-grams offer a simple, but rich representation of conversational data. In order to remove duration effects, we restrict our focus to n-gram frequencies rather than raw n-gram counts. We then scale these frequencies by the log-document-frequency. The $d$th entry into a conversation feature vector is thus given as:

$$x_d = w_d \cdot \frac{c_d}{\sum_{d'} c_{d'}}, \tag{1}$$

where $c_d$ is the count of the $d$th token in the given conversation and $w_d$ is the log-document-frequency scaling:

$$w_d = \log\left(\frac{\text{Num. conversations } d\text{th token appears in}}{\text{Num. conversations}}\right). \tag{2}$$

Empirically, we did not find significant changes in overall accuracy from normalizing for number of spoken words. However, we did find that the log-document-frequency scaling tended to reduce variance within the k-fold cross-validation sweeps.

For all of our experiments we use an SVM classifier, implemented in libSVM [16]. The parameters for SVM were chosen using a 30-fold cross-validated (CV) grid search over the training data. The unigram feature vector contained 25,201 unique tokens while the bigram feature contained 736,033 unique tokens. Experimental results for several basic n-gram features are given in the top section of Table II. First and foremost, it is clear a large and exploitable differences exist between the two classes. For the majority of our experiments we managed to achieve over 80% classification accuracy on the verification data; an

encouraging result given that features were extracted from errorful ASR transcripts. The unigram vector with radial basis function (RBF) kernel performed the best, yielding 87.50% accuracy.

More interesting perhaps is lack of improvement seen from bigrams: 84.76% verification accuracy. This drop in accuracy was consistent with lower accuracies observed during cross validation. We hypothesize two reasons for why moving toward the richer bigram model degraded results. First, ASR transcription error compounds when migrating from unigram to bigram analysis. This makes bigrams much more sensitive to the ASR transcription error. Secondly, the bigram feature set was extremely sparse (736K bigram tokens versus 25K unigram tokens), with a very few bigrams appearing more than a handful of times.

### B. Alternate Features: Stemmed, LIWC, POS

Given the issues with n-gram sparsity it begs the question whether more parsimonious feature sets can achieve similar accuracies. We investigate three lower dimensionality feature sets: 1) stemmed unigrams, 2) a hand-crafted dictionary of salient words from social psychology, and 3) part of speech tags (POS).

Word stemming is used to map families of words down to a morphological base (e.g. jump, jumps, jumped, jumping → jump). Stemming is employed within many machine learning applications were the morphological variants of words are viewed as a source of noise on top of the more important content-conveying word stem. We use an implementation of the Porter stemmer for our experiments [17]. While the stemmer does miss many mappings (e.g. geese, goose) it results in a reduction of the dictionary down to approximately 15,000 word stems.

We investigated a hand-crafted dictionary of salient words, called LIWC, employed in a number of social psychology studies [18]. This dictionary attempts to group words into 64 categories such as pronouns, activity words and positive and negative emotions. The categories have significant overlap and a given word can map to zero or more categories. The clear benefit of LIWC is that the word categories have very clear and pre-labeled meanings. However, since categories are not defined under any sort of mathematical criteria, there are no guarantees the resultant feature will possess any useful discriminative information.

Finally, we examine the use of part-of-speech (POS) tags as a feature set. Though the POS feature is orders of magnitude more compact than the unigrams, our primary goal here is to probe the utility of syntactic versus content driven features. We obtain our POS features by running our ASR transcripts through the Stanford tagger [19]. Stop words were left in place to provide the context necessary for tagging words accurately. We treat the 36 POS tags as a `bag of features', deriving frequency based estimates as per the unigram feature vector.

The bottom section of Table II gives the recognition accuracies for the aforementioned feature sets. While all of the reduced feature sets did posses significant discrimination utility, it is clear varying degrees of performance has been sacrificed. In particular, the degradation caused by unigram stemming was somewhat surprising given that the dictionary was only shrunken from 25K words down to 15K stems. The LIWC feature set saw even further degradation in accuracy, which can perhaps be attributed to its very aggressive word mappings. POS tag features also suffer performance degradation, with accuracy of about 77.7%. Though not as discriminative as the unigrams, it suggests discrimination can be achieved without content-based features and with only coarse features related to speaking style. Furthermore, such POS features can be extracted reliably from noisy ASR transcripts.

To test whether the POS information was complementary to the content-based features, we ran an additional experiment where we created a hybrid unigram-POS feature set. Here, feature tokens were given as a word-POS tuple (denoted by × sign in Table II). Depending on context, a word can be associated with a number of tags. This had a counter-productive effect of increasing the dictionary size to 35,274. Unfortunately the additional information did not seem to provide any benefit over the standard unigram. We also tried simply concatenating the POS feature vector onto both the unigram and LIWC vectors (denoted by + signs in Table II), but similarly, we saw no improvement.

### C. Supervised Feature Selection

Given the mixed results from alternate feature sets, we now investigate the somewhat simpler approach of pruning the unigram dictionary. This dictionary truncation can be undertaken with several criteria in mind, though in this study we limit our scope to frequency based and mutual information (MI) based methods. For frequency based removal, we simply simply drop rare words. Any word fails to appear in a certain number of conversations is dropped. For the MI based method, we calculate the mutual information between word presence and class labels. Words with low mutual information are then selected as candidates for removal. As well as being more principled than the frequency-based truncation, calculation of MI rankings gives us direct insight into the lexicon of the two classes. Table III shows several high MI ranked words for both categories. Here we have shown the top 30 MI ranked words indicative of a business call and 30 words indicative of a social call.

Using either dictionary truncation method, we can achieve various levels of pruning by altering the respective threshold parameters. Dictionary sizes tested range from 250 to the full 25k. The lower limit of 250 was selected to ensure every conversation had at least one in-dictionary word. Figure 1 shows the effect of truncating the unigram dictionary. It is immediately apparent that the unigram dictionary is extremely robust to a wide range of pruning. Not surprisingly, the MI approach gave better verification accuracies – particularly at more aggressive settings. However it is of interest to note that both methods performed remarkably well given their simplistic strategy. A 1000-word dictionary here gave better performance than the much larger stemmed unigram dictionary (15K words), a finding which suggests important information resides within the word morphology. As a final note, we did not observe a monotonic degradation of accuracy when increasing the pruning. Both methods produced a local optima around the 1000-word setting, which appears to be indicative of a trade-off between the negative effect of sparsity and positive effect of model richness.

### D. Unsupervised Feature Selection: Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) was first proposed as a way to generatively model a collection of documents [8]. Here, individual documents are modeled as a mixture of $K$ latent topics, each of which is governed by a multinomial distribution. We are motivated to test LDA models here for three reasons. First it utilizes a `bag of words' model, which may be more robust to ASR error. Second, it is capable of producing parsimonious models without the loss of richness incurred by truncating unigrams. Finally, being a generative model we can leverage the substantial amount of unlabeled data on our corpus.

LDA requires specification of two hyper-parameters. The hyper-parameter $\alpha$ is the Dirichlet prior for the topic multinomial distribution over conversations, while $\beta$ is the Dirichlet prior for the word multinomials over topics. Setting these parameters to small values has the effect of modeling conversations with few topics (in the case of $\alpha$) and modeling topics with few words (in the case of $\beta$).

For our experiments we generate a K-topic LDA model from the 8230 non-verification records. Experimentally, we found best CV results were gained by setting LDA parameters $\alpha$ and $\beta$ to 0.01 and 0.1 respectively. To produce feature vectors, we calculate log-probabilities for the K topics in each conversation. Results for the LDA experiments are shown in Figure 2 for varying numbers of topics. The unsupervised features perform almost as well as unigram features, achieving an accuracy of 86.59% with 30 topic clusters. Though accuracy did drop when the number of clusters were reduced, decent discrimination persisted down to the 2-topic model. This suggests the LDA model was able to approximately learn the our classes in an unsupervised manner. Indeed, when peering into words which constituted the topics, a business/social trend did seem to emerge. Table IV lists words strongly associated with the two topics and clearly the unsupervised clustering appears to have automatically differentiated the business-oriented calls from the rest. On closer examination, we found that most of the probability was distributed in a limited number of words in the business-oriented topic. On the contrary, the probability was more widely distributed among words in the other cluster.

## IV. Error analysis

In this final experiment Section, we perform several error analyses on our data. Since the unigram feature set gave top performance, we use it as the basis for these experiments. Our first error analysis concerns error break-down for individual subjects. Our reason for undertaking this analysis was to ensure our model were not simply learning the lexicon of the more dominant contributors at the expense of the other subjects. Breakdown of the verification error is given in Table V. Despite variation among subjects, clear discrimination was achieved in each case. Encouragingly, homes 4 and 8 achieved very high classification scores despite cumulatively contributing less than 4% of the training corpus.

Our next error analysis concerns error distributions conditioned on conversation lengths. Our hypothesis here was that the longer a (homogeneous) conversation goes on for, the easier it should be to determine its context. To test this, we break the verification data into 5 groups sorted by word counts. Accuracies for each verification sub-group are given in Table VI. Here a fairly clear pattern emerges, with error increasing for the shorter conversations. However, once word counts hit 300–400 words (2–3 minutes of conversation), accuracy tended to plateau.

Our final analysis concerns the separability of the two classes. To do this, we measure the confidence of the classifier against the accuracy it actually achieves. Ideally, a high confidence should be matched with very high accuracy. To measure confidence in the SVM classifier, we use the absolute SVM decision score. As with the previous experiment, we can sort verification examples into sub-groups to individually evaluate accuracy. Results are given in Table VII. Here, results shows clear correlation between classifier confidence and accuracy. The high correlation between confidence level and accuracy allows us to achieve an arbitrary precision, by selectively ignoring `hard to classify' examples. Figure 3 shows the trade-off plot between test set coverage and accuracy. As a final note, the sub-groups of both word counts and confidence levels typically had balanced levels of residential and business conversation. One exception to this was the top 20% confidence group, which was 70% business calls. Further investigation revealed that many of these calls could be traced back to automated customer service conversations – calls which tended to have a very limited and recognizable vocabularies.

## V. Conclusion

In this paper we evaluated the utility of various lexical features in discriminating business-oriented telephone conversations from the others. Similar to results on the topic classification task, we found unigram features to give good discrimination even when they are extracted from noisy ASR transcripts. The classification accuracy is not particularly sensitive to number of features and performance degrades gracefully when the features are pruned using supervised and unsupervised methods. Among the other features explored, we found the performance to be most sensitive to word morphology (e.g. word tense), with word stemming introducing significant reductions in accuracy. Part-of-speech tags possess surprisingly good discrimination ability, given that the POS tags only capture speaking style in coarse manner and is devoid of content. However, its performance is still eclipsed by the unigram feature.

For this paper, the majority of experiments we have presented have relied upon a simple binary classification of the data. While initially an arbitrary choice, this classification was corroborated by the fact that unsupervised two-topic LDA models tended to learn the same boundary. Of course, this begs the question whether topics learned in an unsupervised manner using LDA would separate the conversations along social nature of the calls as well. Our future work will examine such unsupervised methods as well as semi-supervised methods where clusters are seeded, for example, with words related to health or dining out.

## Acknowledgments

This research was supported in part by NIH Grants 1K25AG033723-01A2 and P30 AG024978-05. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NIH. We thank Nicole Larimer for help in collecting the data, Maider Lehr for testing the data collection devices and Katherine Wild for early discussions on this project.

## References

[1]. Bassuk SS, Glass TA, Berkman LF. Social disengagement and incident cognitive decline in community-dwelling elderly persons. Ann Intern Med. 1999; 131(3):165–173. [PubMed: 10428732]

[2]. Taylor, P.; Morin, R.; Parker, K.; Cohn, D.; Wang, W. Growing old in america: Expectations vs. reality. Jun 29. 2009 http://pewsocialtrends.org/files/2010/10/Getting-Old-in-America.pdf

[3]. Shafran I, Mohri M. A comparison of classifiers for detecting emotions from speech. IEEE ICASSP. 2005; 1:341–344.

[4]. Shafran I, Riley M, Mohri M. Voice signatures. IEEE Automatic Speech Regonition and Understanding. 2003:31–36.

[5]. Shriberg E, Stolcke A, Jurafsky D, Coccaro N, Meteer M, Bates R, Taylor P, Ries K, Martin R, van Ess-Dykema C. Can prosody aid the automatic classification of dialog acts in conversational speech? Language and Speech. Jul-Dec;1998 41(3–4):443–492. [Online]. Available: http://las.sagepub.com/content/41/3-4/443.abstract. [PubMed: 10746366]

[6]. Stolcke A, Ries K, Coccaro N, Shriberg E, Bates R, Jurafsky D, Taylor P, Martin R, Ess-Dykema CV, Meteer M. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics. 2000; 26(3):339–373. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/089120100561737.

[7]. Jurafsky D, Bates R, Coccaro N, Martin R, Meteer M, Ries K, Shriberg E, Stolcke A, Taylor P, Van Ess-Dykema C. Automatic detection of discourse structure for speech recognition and understanding. Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on. Dec.1997 :88–95.

[8]. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J. Mach. Learn. Res. Mar.2003 3:993–1022. [Online]. Available: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.

[9]. Shriberg E, Stolcke A, Hakkani-Tr D, Tr G. Prosody-based automatic segmentation of speech into sentences and topics. Speech Communication. 2000; 32(1–2):127–154. accessing Information in Spoken Audio. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639300000285.

[10]. Wallach, HM. Topic modeling: beyond bag-of-words. Proceedings of the 23rd international conference on Machine learning, ser. ICML '06; New York, NY, USA: ACM; 2006. p. 977-984. [Online]. Available: http://doi.acm.org/10.1145/1143844.1143967

[11]. Mehl MR, Vazire S, Holleran SE, Clark CS. Eavesdropping on happiness. Psychological Science. 2010 [Online]. Available: http://pss.sagepub.com/content/early/2010/02/17/0956797610362675.short.

[12]. Godfrey, J.; Holliman, E.; McDaniel, J. SWITCHBOARD: Telephone speech corpus for research and development. IEEE International Conference on Acoustics, Speech, and Signal Processing; 1992. p. 517-520.

[13]. McDonough J, Ng K, Jeanrenaud P, Gish H, Rohlicek J. Approaches to topic identification on the switchboard corpus. Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on. Apr.1994 i:I/385–I/388. vol.1.

[14]. Gillick L, Baker J, Baker J, Bridle J, Hunt M, Ito Y, Lowe S, Orloff J, Peskin B, Roth R, Scattone F. Application of large vocabulary continuous speech recognition to topic and speaker identification using telephone speech. Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on. Apr.1993 2:471–474. vol.2.

[15]. Soltau, H.; Kingsbury, B.; Mangu, L.; Povey, D.; Saon, G.; Zweig, G. The IBM 2004 conversational telephony system for rich transcription. IEEE International Conference on Acoustics, Speech, and Signal Processing; 2005. p. 205-208.

[16]. Chang C-C, Lin C-J. Libsvm : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2

[17]. van Rijsbergen C, Robertson S, Porter M. New models in probabilistic information retrieval. British Library Research and Development Report. 1980; (5587) [Online]. Available: http://tartarus.org/ martin/PorterStemmer/.

[18]. Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology. 2003; 54(1):547–577.

[19]. Toutanova, K.; Manning, CD. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora; 2000. p. 63-70.

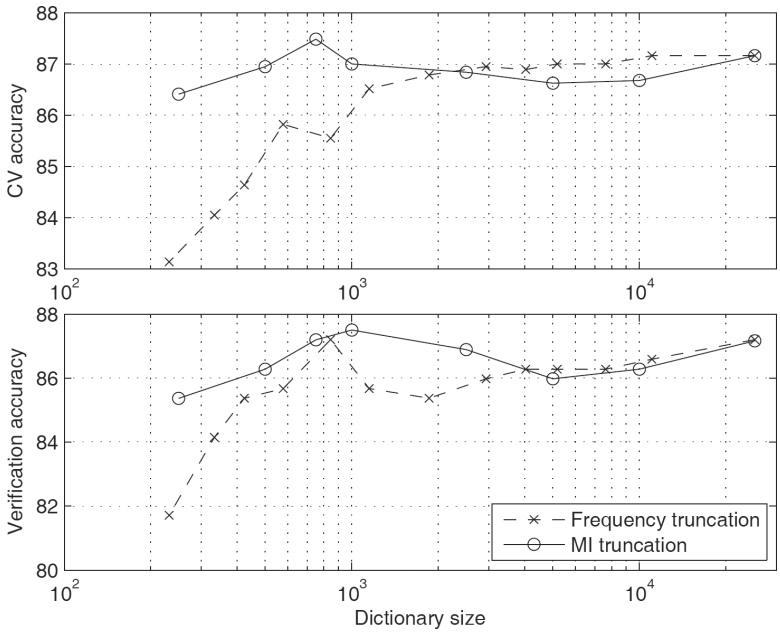


**Fig. 1.**

Recognition accuracy under the effect of unigram dictionary pruning. Top plot shows cross validation accuracies while bottom plot shows verification set accuracies.

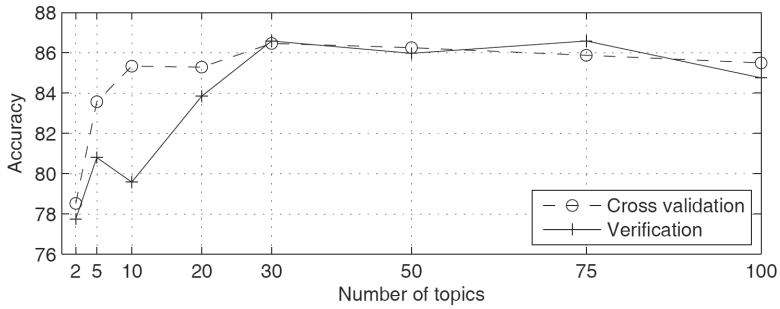


**Fig. 2.**
Classification using LDA topic log-probabilities.

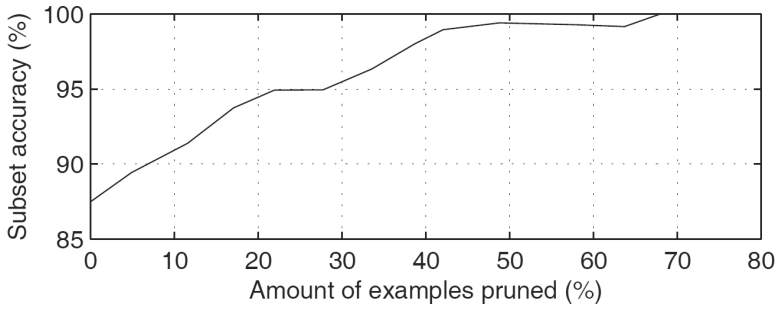


**Fig. 3.**

Accuracy of the unigram classifier with confidence pruning. Confidence is based on the SVM output score. Testing examples with the lowest absolute scores are pruned first.

**TABLE I**

Contribution of each house to the telephony corpus. Houses marked with an asterisk indicate a phone shared by a couple.

| Home ID | Training (res) Instances | Training (biz) Instances | Training (all) Instances | Testing (res) Instances | Testing (biz) Instances | Testing (all) Instances |
|---|---|---|---|---|---|---|
| 1 | 15 | 18 | 33 | 3 | 5 | 8 |
| 2* | 188 | 338 | 526 | 40 | 63 | 103 |
| 3 | 135 | 160 | 295 | 21 | 21 | 42 |
| 4 | 7 | 27 | 34 | 1 | 5 | 6 |
| 5 | 113 | 67 | 180 | 18 | 9 | 27 |
| 6* | 312 | 129 | 441 | 54 | 20 | 74 |
| 7 | 88 | 53 | 141 | 17 | 8 | 25 |
| 8 | 73 | 139 | 212 | 10 | 33 | 43 |
| TOTAL | 931 | 931 | 1862 | 164 | 164 | 328 |

*Proc IEEE Workshop Autom Speech Recognit Underst*. Author manuscript; available in PMC 2012 June 27.

**TABLE II**

Classification accuracies for basic word features. Accuracies are given for the hold-out verification testing data.

| Feature setting | Accuracy (%) |
|---|---|
| Linear unigram | 82.93 |
| RBF unigram | 87.50 |
| RBF bigram | 84.76 |
| RBF stemmed unigram | 83.23 |
| RBF LIWC | 78.66 |
| RBF POS | 77.71 |
| RBF POS+LIWC | 78.96 |
| RBF POS+unigram | 86.89 |
| RBF POS×unigram | 86.58 |

**TABLE III**

Selection of words with high class label mutual information.

| **Business oriented** | **Social oriented** |
|---|---|
| Press, thank, calling, information, service, customer, number, quality, please, pressed, representative, account, zero, seven, monitored, transferred, nine, six, transfer, services. | Hi, dinner, she's, high, Brandon, home, dad, night, everybody, doing, tonight, later, hello, mom, anyway, bad, nice, sleep, tomorrow, house. |

## TABLE IV

Binary topics for latent Dirichlet allocation.

| Topic 1 | Topic 2 |
|---|---|
| Invalid, helpline, eligibility, transactions, promotional, representative, mastercard, touchtone, activation, nominating, receiver, voicemail, digit, representatives, Chrysler, ballots, staggering, refills, resented, classics, metro, represented, administer, transfers, reselling, recommendations, explanation, floral, exclusive, submit. | Adorable, aeroplanes, Arlene, Astoria, baked, biscuits, bitches, blisters, bluegrass, bracelet, brains, bushes, calorie, casinos, Char-lene, cheeses, chit, Chris, clam, clientele, cock, cookie, copying, crab, Davenport, debating, dementia, dictionary, dime, Disneyland, eek, Eileen, fascinated, follies, fry, gained |

## TABLE V

Verification error breakdown by subject.

| Home | Records | Accuracy |
|---|---|---|
| 1 | 8 | 87.50 |
| 2 | 103 | 84.47 |
| 3 | 42 | 80.95 |
| 4 | 6 | 100.00 |
| 5 | 27 | 77.00 |
| 6 | 74 | 94.59 |
| 7 | 25 | 88.00 |
| 8 | 43 | 90.70 |

**TABLE VI**

Verification error breakdown versus conversation length.

| Word count percentile | Word count range | Acc. | Res / Biz split | Res acc. | Biz acc. |
|---|---|---|---|---|---|
| 0–20 | 30–87 | 75.76 | 62.12 / 37.88 | 82.93 | 64.00 |
| 20–40 | 88–167 | 83.33 | 48.48 / 51.52 | 75.00 | 91.18 |
| 40–60 | 168–295 | 90.91 | 39.39 / 60.61 | 96.15 | 37.50 |
| 60–80 | 296–740 | 93.94 | 40.91 / 59.09 | 96.30 | 92.31 |
| 80–100 | 741+ | 93.75 | 59.38 / 40.62 | 100.00 | 84.62 |



*Proc IEEE Workshop Autom Speech Recognit Underst*. Author manuscript; available in PMC 2012 June 27.

**TABLE VII**

Verification error breakdown versus SVM margin confidence.

| Margin percentile | Margin range | Acc. | Res / Biz split | Res acc | Biz ace |
|---|---|---|---|---|---|
| 0–20 | 0–0.345 | 61.54 | 47.69 / 52.31 | 70.97 | 52.94 |
| 20–40 | 0.345–0.74 | 80.60 | 55.22 / 44.78 | 81.08 | 80.00 |
| 40–60 | 0.74–1.06 | 96.97 | 65.15 / 34.85 | 97.67 | 95.65 |
| 60–80 | 1.06–1.48 | 98.48 | 51.52 / 48.48 | 100.00 | 96.88 |
| 80–100 | 1.48+ | 100.00 | 29.69 / 70.31 | 100.00 | 100.00 |



*Proc IEEE Workshop Autom Speech Recognit Underst*. Author manuscript; available in PMC 2012 June 27.