Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a preprint version which may differ from the publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/122950

Please be advised that this information was generated on 2017-12-05 and may be subject to change.

Combining textual and non-textual features for e-mail importance estimation

Maya Sappelli [a] Suzan Verberne [b] Wessel Kraaij [a]

[a] *TNO and Radboud University Nijmegen*
[b] *Radboud University Nijmegen*

Abstract

In this work, we present a binary classification problem in which we aim to identify those email messages that the receiver will reply to. The future goal is to develop a tool that informs a knowledge worker which emails are likely to need a reply. The Enron corpus was used to extract training examples. We analysed the word n-grams that characterize the messages that the receiver replies to. Additionally, we compare a Naive Bayes classifier to a decision tree classifier in the task of distinguishing replied from non-replied e-mails. We found that textual features are well-suited for obtaining high accuracy. However, there are interesting differences between recall and precision for the various feature selections.

1 Introduction

In the COMMIT project SWELL (smart reasoning for well-being at work and at home[1]) we aim to support knowledge workers in their daily life. In the at work scenario one of the objectives is to prevent negative stress, either by coaching the user on his work style or by signalling stressed behaviour [12]. Another objective is to filter irrelevant information to preserve the user's work flow. A large source of incoming information for knowledge workers is e-mail.

For this latter objective there are three things that are important. First we need to know what the user is doing to determine which incoming messages are relevant for his current work, and whether presenting the user with the message is disturbing. We define this as recognizing the user's context. Second, we need to decide which incoming messages are important enough to present it to the user regardless of what he is doing. This aspect is important to make the user feel in control, i.e. that he does not feel like he is missing information. Third, it is important to understand why an incoming message is important or relevant so this can be used as feedback to the user (i.e. transparency).

In this paper we focus on assessing the likeliness that a receiver will reply to a message. We believe that replying to a message is a good indicator that a user finds this message important, otherwise he would have ignored it. We stress that the likeliness of reply is not the only factor determining message importance, but it is a start. This work is meant as a first step towards developing an e-mail client that helps to protect the user's work flow. Existing literature on the topic of reply prediction (section 2) focuses on features such as the number of question marks and the number of receivers. We aim to investigate the influence of the textual content of the message on the likeliness that a receiver will reply to the message. This can also be used to make it transparent to the user why a classifier beliefs that the user needs to reply to a message. To this end, we train classifiers with various feature sets and compare their results.

[1] `www.swell-project.net`

2 Related Work

This section presents an overview of the literature related to reply prediction. First, we present some general work on email responsiveness. After that we present some previous attempts to manual or automatic prediction of whether an e-mail message is going to be replied to.

Tyler et al.[15] conducted a study to email responsiveness to understand what information is conveyed by the timing of email responses. They used interviews and observations to explore the user's perceptions of their own responsiveness and their expectation of responses from other users. They distinguish response expectation from breakdown perception. The former is the implicit time the sender gives to the recipient to respond, which is usually based on the time it took in previous interactions with the recipient. The latter is the initiation of a follow-up action, that occurs when the response expectation time has ended, which is dependent on the recipient, the recipient's location, the topic urgency and whether a voice mail was sent. These findings suggest that the social context of a message might be more important than the contents of the message.

In a survey study with 124 participants, Dabbish et al [5, 6] investigated what characteristics of email messages predict user actions on messages. The authors present a model of reply probability based on the outcomes of this survey. Important factors were the importance of the message, number of recipients, sender characteristics and the nature of the message content. Sender characteristics seemed to have the greatest effect. They did not find an effect of communication frequency on reply probability and suggest that this may be due to the availability of other communication channels that reduce the necessity for email response. The perception of message importance was influenced by (1) communication frequency in combination with the status of the sender,(2) whether the message contained a status update, and (3) whether the message was a scheduling event.

There has been several attempts to automatic reply prediction. Dredze et al.[8] developed a logistic regression predictor that indicates whether email messages necessitate a reply. Their predictor was evaluated on the spam-free inbox and sent-mail folders of two graduate students. Features used were word identity, message length, whether the message contained the mentioning of a date and time, whether the recipient was directly addressed, whether it contained a question and who the recipients or sender was. ROC curves of the trained logistic regression model revealed that to achieve 80 % true positives (message predicted to receive a reply that were actually replied to) there were 50% false positives (message predicted to receive a reply that were not replied to)

In later research Drezde et al.[9] used a rule based system to predict reply labels (needs reply, does not need reply). In this system they used relational features that rely on a user profile which included the number of sent and received emails from and to each user as well as the user's address book, a supervisor-role indication, email address and domain. Document-specific features were the presence of question marks, request indicators such as question words (weighted using tf-idf scores), presence of attachment, document length, salutations, and the time of day. The system was tested on 2,391 manually labelled emails, coming from 4 students. On average it obtained a precision of 0.73 and recall of 0.64.

In larger scale research using the Enron corpus[11, 3], Deepak et al. [7] and On et al. [14] investigate the responsiveness and engagingness of users. Their models are based on the number of received replies and the number of sent replies as well as the time it takes to reply. They do not take any content into account.

Ayodele et al.[2] use the Enron corpus to develop and evaluate a manual rule-based reply prediction method. They use largely the same features as Dredze et al.[9] In a second approach they use only the presence of certain words, salutations, question marks, dates or month names and AM or PM. For both approaches the authors report to have very high accuracies of 100% and 98%. These results are unrealistically high because the e-mails are evaluated manually by human reviewers using the described rules.

In more general research, Aberdeen et al.[1] try to predict the probability that the user will interact with an email (i.e. open, read, forward or reply) within a certain time span. They use a linear regression model and a form of transfer learning to determine a ranking of the interaction likeliness. A threshold determines which messages are indicated as important.They have used social features (based on interaction with recipient), content features (headers and recent terms that are highly correlated with actions on a message), thread features (interaction of the user with a thread) and label features (labels applied to the message using filters).

They obtained an accuracy of 80%. Their work is the basis for the Google Priority Inbox.

3 Method

The goal of this experiment is to assess whether textual content features have added value when it comes to predicting whether a message will receive a reply or not. For that purpose we select textual features using various feature selection methods (described in Section 3.2). We analyse the selected features on their transparency (i.e. how easy are they to interpret?) and evaluate their effectiveness in a classification experiment (Section 3.3). We start this section with the description of how we obtained our labelled dataset.

3.1 Extracting threads from the Enron corpus

To obtain a labelled dataset, we constructed threads from the Enron corpus to determine which message had received a reply and which not. We have used the August 2009 version of the corpus without file attachments to have a fair comparison the existing literature. We have taken a tree based approach [17] for extracting the threads from Enron using the algorithm suggested by [7]. From these threads we derived which messages were replies by matching the subject lines and including the prefix "RE:" (case-insensitive). For each reply message we found the corresponding original message (i.e. the message that was replied to) by selecting the message with the same thread id, of which the sender was a receiver of the reply and which was sent before the reply. In the rare case that there were multiple options, we chose the message that was closest in time to the reply. Out of the 252,759 messages in the Enron corpus, we found 3,492 messages that have received a reply and 166,572 message that have not received a reply. We do not take into account messages that are forwards or replies on replies.

3.2 Feature Selection

We have used three different methods for analysing the influence of the textual content of the messages. The first measure is χ^2 [18], which measures the dependence between a term and a class. We are looking for the terms with a high dependency on the replied-class (i.e. a high χ^2 score).

$$\chi^2(t,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{1}$$

where A is the number of *replied* messages that contain term t, B is the number of *non-replied* messages that contain t, C is the number of *replied* messages that do not contain term t and finally, D is the number of *non-replied* messages that do not have t. N is the total number of messages in the set.

The second method we used is point-wise Kullback-Leibler divergence [13], as suggested by [4]. This measure determines a value for each term which indicates how good that term is for distinguishing the set of replied messages from the set of non-replied messages.

$$KLdiv(t,p) = P(t|p)log\frac{P(t|p)}{P(t|n)} \tag{2}$$

where $P(t|p)$ is A and $P(t|n)$ is B.

The third and final method is based on linguistic profiling as proposed by [16]. It compares the normalized average term frequency of a term in the positive set (replied messages) to its average in the negative set (non-replied messages). Rather than using the proposed classification method, we use linguistic profiling for term selection.

$$LP(t) = \mu(t,p) - \mu t, n \tag{3}$$

where $\mu(t,p)$ denotes the normalized average frequency of term t in the set of replied messages and $\mu(t,n)$ denotes the normalized average frequency of term t in the set of non-replied messages.

With all three methods, we extracted the most important terms from the example set. As terms, we considered all word n-grams with $n \in 1,2,3$, and we use the number of occurrences of each n-gram to represent a message.

3.3 Classification

In a classification experiment, we compare the effectiveness of the feature selection methods from the previous section to the effectiveness of the features described in literature. These features (referred to as non-textual features) are: (1) number of receivers in the fields TO, CC and BCC respectively, (2) number of question marks, (3) number of previously received replies from recipient (4) likeliness of interaction with receiver (5) message length (6) occurrence of each of the question words *what, when, where, which, who, whom, whose, why* and *how* weighted with tf-idf . For each of the textual feature selection methods, selections of 10, 50, 100, 500, 1000 and 5000 features were compared.

The original distribution contains 97% negative examples (non-replied e-mails), which is very imbalanced. Therefore, we first rebalance our data by selecting two random negative examples for each positive example in our data. We split our data into 90% train (10476 examples) and 10% test (1167 examples). All examples in the test set have later dates than the examples in the train set to prevent leaking of future information in the training data. We used a Naive Bayes classifier and a J48 decision tree classifier from the WEKA toolkit [10], with their default settings. Typically decision tree works well for non-text features and Naive Bayes is well-suited for textual features. The WEKA resample filter is used to balance the data uniformly by oversampling the positive examples. The reason for first under balancing the negative examples is to prevent a too extreme oversampling of the positive examples. The results were evaluated on the fixed unbalanced test set.

4 Results

4.1 Feature Analysis

Table 1: Top 10 n-grams that indicate that a message will receive a reply

$\chi^2(t,c)$	$KLdiv(t,p)$	$LP(t)$
me	i	fyi
keep	we	i
i	me	me
2001	you	we
information	have	you
decisions	know	know
one of	let	have
let	please	http
news	let me	let
receive a	me know	2001

The top 50 n-grams, of which 10 are presented in table 1, of each of the three feature selection methods were manually analysed. Both the point-wise Kullback Leibler and the Linguistic Profiling method indicate the importance of the personal pronouns *I*, *we* and *you*. These pronouns may indicate that the receiver is addressed personally. All methods also seem to indicate the occurrence of the phrase "please let me know" which suggests that the sender expects an action from the receiver. Worth noting is that Linguistic Profiling indicates the importance of the term "fyi". Even though this does not seem intuitive, inspection of messages reveals that "fyi" messages often receive a "thank you" reply. The terms selected by the χ^2 measure seem to be less easy to interpret. They may refer to more specific situations. Overall, first analysis suggests that of the top 50 terms point-wise Kullback Leibler term selection is the easiest to interpret and the least sensitive to noise.

Table 2: Classification results for the optimal number of features. Reported precision and recall are for the "will reply" class only. Best results are indicated in bold face. BOW refers to a full bag of words frequency model

		Naive Bayes		
Feature Type	# Features	Accuracy	Precision	Recall
non-Text		42.6%	0.358	**0.912**
$\chi^2(t,c)$	1000	59.0%	0.43	0.709
$KLdiv(t,p)$	10	70.8%	**0.635**	0.291
$LP(t)$	50	**72.0**%	0.586	0.544
BOW	117400	58.7%	0.417	0.611
		Decision Tree		
Feature Type	# Features	Accuracy	Precision	Recall
non-Text		69.9%	0.581	0.351
$\chi^2(t,c)$	10	66.9%	0.502	0.557
$KLdiv(t,p)$	500	65.7%	0.475	0.291
$LP(t)$	50	64.1%	0.441	0.296
BOW	117400	62.2%	0.432	0.436

4.2 Classification

Table 2 shows the classification results for the optimal number of features with the various feature selection methods and the two classification approaches. The reported precision and recall are for the "will reply" class only.[2]

When we look at the Naive Bayes results in table 2 we see that if we select as little as 50 features from the LP measure we have a reasonable accuracy (72.04%). The classifier with only non-textual features, performs much worse and shows an accuracy of 42.62%. Interestingly its recall for the positive class is very high: it recognizes more than 90% of the emails that received a reply.

When we look at the results for the decision tree, we see that the classifier with non-textual features performs better than with Naive Bayes (69.98%), while the runs on only textual features selected by $\chi^2(t,c)$, $KLdiv(t,p)$ and $LP(t)$ all give an accuracy around 65%. Interestingly, $\chi^2(t,c)$ performs a lot better than in the Naive Bayes classifier, while $KLdiv(t,p)$ and $LP(t)$ perform worse. We only found very small differences in classification performance when we vary the number of selected features.

Combined classifiers that were trained on the combinations of textual and non-textual features performed approximately as good as the best classifier of the two. It is interesting to notice that the performance of the feature selection method $\chi^2(t,c)$ is so different with the two classifiers. $\chi^2(t,c)$ is often used as a feature selection method for text classification, especially in Naive Bayes, while this experiment suggests that point-wise Kullback-Leiber divergence and Linguistic Profiling might be better feature selectors.

5 Conclusion

In the current work we found that after first analysis of three feature selection methods for reply prediction, point-wise Kullback Leibler divergence seems a useful measure to select interpretable terms that are representative of a class. Linguistic Profiling seems suitable as well but seems to contain a little more noise.

Using a Naive Bayes classifier we only need as little as 50 terms selected by linguistic profiling to achieve a reasonable accuracy (72.04%). This is even better than our baseline results with non-text features with a decision tree (69.98%), but only slightly. On the other hand, we obtained the highest recall with non-text features.

Concluding, we can predict with reasonable accuracy which e-mails will be replied to. Although, 72% success might not be accurate enough to be used as a stand-alone application, we can use it as an indication

[2]Since this task does not include ranking of messages, some evaluation metrics such as 11-pt interpolated average precision and precision@k could not be applied

of how important that message is. However, whether a message will be replied to is likely not the only determinant of message importance, so future work may include other methods for estimating message importance.

Additionally, transparency is an important concept in SWELL, and we think that it is important to find a good balance between precision and recall, so that the user has trust in the system (i.e. does not feel like important messages are missed), but also understands why some indications are given, and does not require too much additional feedback. Given the results of our experiment it seems important to find a method that combines a classifier with high recall such as Naive Bayes with non-text features, and a classifier with high precision such as Naive Bayes with features selected by Kullback-Leibler divergence.

6 Acknowledgements

This publication was supported by the Dutch national program COMMIT (project P7 SWELL). Additionally, we thank Micha Hulsbosch for his help in optimizing the thread extraction algorithm.

References

[1] D. Aberdeen, O. Pacovsky, and A. Slater. The learning behind gmail priority inbox. In *LCCC: NIPS 2010 Workshop on Learning on Cores, Clusters and Clouds*, 2010.

[2] T. Ayodele and S. Zhou. Applying machine learning techniques for e-mail management: solution with intelligent e-mail reply prediction. *Journal of Engineering and Technology Research*, 1(7):143–151, 2009.

[3] R. Bekkerman. Automatic categorization of email into folders: Benchmark experiments on Enron and SRI corpora. *Computer Science Department Faculty Publication Series*, page 218, 2004.

[4] C. Carpineto, R. de Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19(1):1–27, Jan. 2001.

[5] L. Dabbish, R. Kraut, S. Fussell, and S. Kiesler. To reply or not to reply: Predicting action on an email message. In *ACM 2004 Conference*. Citeseer, 2004.

[6] L. Dabbish, R. Kraut, S. Fussell, and S. Kiesler. Understanding email use: predicting action on a message. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 691–700. ACM, 2005.

[7] P. Deepak, D. Garg, and V. Varshney. Analysis of Enron email threads and quantification of employee responsiveness. In *Workshop on Text Mining and Link Analysis (TextLink 2007)*, 2007.

[8] M. Dredze, J. Blitzer, and F. Pereira. Reply expectation prediction for email management. In *The Second Conference on Email and Anti-Spam (CEAS), Stanford, CA*, 2005.

[9] M. Dredze, T. Brooks, J. Carroll, J. Magarick, J. Blitzer, and F. Pereira. Intelligent email: Reply and attachment prediction. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 321–324. ACM, 2008.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[11] B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. *Machine Learning: ECML 2004*, pages 217–226, 2004.

[12] S. Koldijk, M. Neerincx, and W. Kraaij. Unobtrusively measuring stress and workload of knowledge workers. In *Proceedings of Measuring Behavior*, 2012.

[13] S. Kullback and R. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[14] B. On, E. Lim, J. Jiang, A. Purandare, and L. Teow. Mining interaction behaviors for email reply order prediction. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, pages 306–310. IEEE, 2010.

[15] J. Tyler and J. Tang. When can I expect an email response? a study of rhythms in email usage. In *Proceedings of the eighth conference on European Conference on Computer Supported Cooperative Work*, pages 239–258. Kluwer Academic Publishers, 2003.

[16] H. Van Halteren. Linguistic profiling for author recognition and verification. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 199. Association for Computational Linguistics, 2004.

[17] G. Venolia and C. Neustaedter. Understanding sequence and reply relationships within email conversations: a mixed-model visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 361–368. ACM, 2003.

[18] Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, pages 412–420. Morgan Kaufmann Publishers, Inc., 1997.