Is this Paragraph Important?

Classifying ‘Terms of Service’ Components

Data Mining Final Report

Matt Hawthorn, Jon Lazenby, Hope McIntyre, Katherine Schinkel

Problem Description

- **What is the problem?**

Currently, Terms of Service (ToS) agreements are long and difficult to understand. Most, if not all, internet services either explicitly require that a user agrees to the ToS or assume that a user agrees by just using the service. It is time-prohibitive for a particular user to read and understand the ToS agreements that apply to all online services that they use. Additionally, considering the length and complexity of these documents, it is unlikely that all users fully understand them, leaving the user vulnerable to privacy and data rights infringement.

- **Why is it important?**

ToS agreements are legally binding documents, and it is important for users of an online service to be aware of and understand the legal parameters that outline their rights, particularly related to what a company can do with their data. Typically, a user may trust that a company will respect their privacy, without fully understanding the disclaimers that are outlined in the ToS agreements. This allows companies to potentially take advantage of a user’s negligence, if they assume that only a small minority of users will actually read the document. Providing this legal information in a digestible manner to users would increase accountability for companies and increase the likelihood that a user’s rights are protected.

- **Who cares about it?**

All users of online services are affected by ToS agreements everyday. These services include Facebook, Google, Amazon, etc., each with millions of users.

- **Why does it remain unsolved?**

ToS agreements have been reviewed manually and summarized; however, there are limitations with this methodology. Manual reading requires substantial resources and is difficult to scale to the growing size of online services and their ToS agreements. These documents are also dense, long, and generally difficult to understand, creating an opportunity for the use of data mining.

Objectives

- **What are you proposing to do about the problem?**

Our plan is to construct a model that accurately identifies important paragraphs within a ToS document. Rather than sifting through the clutter, users would be able to execute this model and retrieve only the paragraphs that are vital to understand. This would ultimately allow users to efficiently navigate a ToS document for critical paragraphs, thereby reducing time and energy.

- **How will you measure the success of your work?**

We are treating this as a binary classification problem. The metrics that we considered when measuring success of our model were precision, recall, accuracy, and F-score. We were particularly interested in maximizing recall because we wanted to minimize false negatives; we don't want our end users to miss truly important paragraphs. Precision is also important because we want to filter the paragraphs of each document down to a manageably readable size. Thus, F is an appropriate measure which effectively represents both recall and precision by taking their harmonic mean:

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

Related Work

- **What have others done to address this problem?**

There is limited research in this space. Researchers at the Georgia Institute of Technology manually "reviewed 30 popular social networking and creative community sites that encourage people to share material, examining the rights to use work that were claimed in the sites' terms of service agreements" [1]. They read each ToS and created a "cod[e] for each license and right mentioned, … whether the site included plain language explanations of copyright terms, [and] whether there were any explicit waivers of rights" [2]. Their research revealed that the ToS within their sample were on average written at a college sophomore reading level [2]. With this in mind and the fact that the "average adult reading speed [is] 250 words per minute, [each ToS] would take almost 8 hours to read" [2]. This research was conducted with the primary goal of better understanding how misinformed the public is on copyright issues within ToS.

A number of articles offer advice to users on how to interpret ToS, but these are designed to guide a user through the review process. These include articles like "Didn't Read Those Terms of Service? Here's What You Agreed to Give Up" by the New York Times Bits blog, which describes the results from the GA Tech study described above in layman's terms, and "How to Quickly Read a Terms of Service" by LifeHacker [1,3].

[1] Singer, Natasha. "Didn't Read Those Terms of Service? Here's What You Agreed to Give Up." New York Times Bits Blog. New York Times, 28 Apr. 2014. Web. 17 Nov. 2015.

[2] Fiesler, Casey, Jessica L. Feuston, and Amy S. Bruckman. "Understanding Copyright Law in Online Creative Communities."

[3] Klosowski, Thorin. "How to Quickly Read a Terms of Service." Lifehacker. Lifehacker, 12 Mar. 12. Web. 17 Nov. 2015.

A website, *ToS Did Not Read* (TOSDR), has been working to help users quickly understand ToS [4]. The non-profit project extracts key pieces from the ToS from sites like Youtube and Facebook. It outlines key pieces such as “Facebook automatically shares your data with many other services” and gives a rating of “Thumbs Up” or “Thumbs Down” for that component of the ToS. It then attempts to provide a class rating (A through E) for how well the ToS respects and protects user rights. The extraction of content and classification is done manually and the information is only available for a limited subset of ToS agreements.

- **What are you doing that is similar to past work?**

Our work is similar to the aforementioned work in that we hope to help users better understand ToS in less time. Like the work done by GA Tech and TOSDR, we are trying to do this by directing the user’s attention to the key components of the ToS.

- **Are there commercial products that accomplish what you are trying to do? What are their characteristics? Where are their gaps?**

There are no commercial products trying to automate the task of pre-processing ToS agreements for easier review. Individual companies are working to make their ToS’s shorter or less technical, but there is not, known from our research, a company trying to make a tool that can be applied to all ToS to make them easier to understand.

- **What about your work is novel? What gaps does it fill?**

As outlined above, utilizing machine learning to process ToS agreements for easier comprehension is entirely novel. We are filling a gap by making a process that can be applied to any ToS, even those that are new or recently modified, without manual intervention. Additionally, this model could fill a gap in Text Mining classification modeling in the legal document space which also, from our research, is limited.

Approach

- **What data did you use? How did you preprocess it?**

Using the python library BeautifulSoup, we scraped the ToS of 13 companies: Yahoo, Google, GitHub, Wikipedia, Amazon, SoundCloud, Twitter, Cloudant, Instagram, Netflix, Facebook, Youtube, and iCloud. By using three different javascript paragraph separators, we divided each ToS into paragraphs and placed each one into a row of a pandas dataframe, together with its company name. A few paragraphs were read in twice, which we manually corrected for by removing the redundant entries. Additionally, multiple paragraphs from Wikipedia’s ToS were not correctly read in, and thus were not included in our modeling. Our final dataset was left with 12 ToS agreements. Future iterations of this script could correct for these issues; a final data product would need to be able to handle incoming ToS’s with unpredictable html formatting.

[4] "About - Terms of Service; Didn't Read." Terms of Service; Didn't Read. Web. 17 Nov. 2015.

Once we had a dataframe of paragraphs, we performed feature engineering to add variables to our dataframe. These paragraph features include: length of paragraph (by character count), count of spaces, count of capitalized characters, location of each paragraph in the full ToS (to indicate whether the paragraph is near the beginning or end, etc.), count of words, count of quotation marks, count of sentences, count of parentheses, and average word length. Additionally, we added binary flags to indicate the presence of the terms “arbitration”, “third party”, and “waiver”, which literature indicated could be correlated with importance. Once this step of feature engineering was complete, we exported this dataframe as a csv.

To generate a response variable, we manually read each paragraph and annotated whether or not we thought the paragraph was important. The criteria for importance was determined by gut-feeling (i.e. “do I care?”) as well as recommendations from articles discussed in the related works section above [1,2,3]. Particular ideas that team members “cared” about were shared with other team members to improve consistency in the importance annotation. Examples include whether the company can change their ToS, has the right to delete a user’s account at anytime, or limits a user’s right to sue.

Although we investigated legal notions of importance prior to annotation, the potential existed for an appreciable degree of subjectivity in our judgements of the documents. Thus we needed a metric by which to compare raters and determine whether their annotations were consistent. In order to compare every pair of raters on an equal basis in terms of agreement (with the minimum amount of manual annotation), we followed common practices of experimental design and assigned raters to the 12 annotation tasks according to a block design with parameters (12, 4, 2, 6). That is to say, there were 4 raters, where each rater read the ToS’s of 6 companies and every ToS was read by 2 raters. Agreement for each pair of raters was quantified by its Cohen’s kappa value. To account for conflicts in classification between raters, we used a response of ‘1’ only for paragraphs where both raters agreed. While this methodology did not follow the traditional Cohen’s process for seeking interrater consensus, we felt this was the best approach to resolve conflicts considering time and resource constraints.

- **What analyses did you perform?**

The first analysis we performed was the computation of Cohen’s kappa. This was done to ensure that the agreement between raters was better than random chance and could be used as a response variable. Cohen’s kappa is defined as the ratio of two differences: the difference between actual agreement and expected agreement due to chance, and the difference between the maximum possible agreement and the expected agreement due to chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

Below is a table of the values of kappa that we computed for each pair of raters:

Rater 1	Rater 2	Cohen's kappa
JL	HM	0.64
JL	MH	0.80
JL	KS	0.71
HM	MH	0.49
HM	KS	0.70
MH	KS	0.48

According to the recommendations of Landis & Koch [5], these values indicate "substantial" agreement for 4 pairs of raters, and "moderate" agreement for the other 2. This was a positive result, though there is still room for future improvement, for instance by employing expert raters or by applying clearer guidelines for importance.

Next, we reviewed the results of a Latent Dirichlet allocation (LDA) topic model on the entire corpus. This analysis was done to determine the effectiveness/clarity of the resulting topics. An analysis of the fitted topics allowed us to conclude that at k = 50, topics were understandable and distinct. We also concluded that stemming the corpus gave rise to more unique topic terms. Additionally, initial experiments using logistic regression showed that the stemmed topics tended to have higher predictive power than the non-stemmed counterparts. While we would have preferred to tune each LDA parameter systematically by its predictive power, LDA is computationally costly. Thus, for this initial proof-of-concept we settled on 50 topics, use of stemming, and an automatically tuned alpha.

We next assessed the viability of a logistic regression model using the engineered features, LDA topic proportions, and a random split of paragraphs into testing and training sets. The evaluation metrics from this type of division were not strong. We hypothesized that the model may perform better with a training/testing dataset divided by company. This has the advantage of better representing the real-life implementation the model. In the actual use-case, a user would present the model with a new complete ToS to classify, rather than a random selection of paragraphs from various ToS's.

Adopting this validation approach with a held-out test set of companies and our chosen response variable (a value of 1 where both raters agreed on the importance of a given paragraph, 0 otherwise), we explored different threshold values for positive prediction. It was found that the optimal value of the threshold was quite variable, possibly owing to the relatively small size of our dataset or imperfections in rater agreement. Rather than tune the threshold as

[5] Landis, J Richard, and Gary G Koch. "The measurement of observer agreement for categorical data." *biometrics* (1977): 159-174.

a part of the validation phase, we adopted a fairly conservative approach and set this value to 0.3. Ideally, a final application would allow an end user to set this value themselves, determining how much they are willing to read.

- **What models did you build?**

Following the exploratory phase detailed above, we settled on employing a logistic regression model using the logical AND of the two rater annotations as the response, and the engineered paragraph features and LDA topic proportions as regressors.

- **What evaluation setup did you use?**

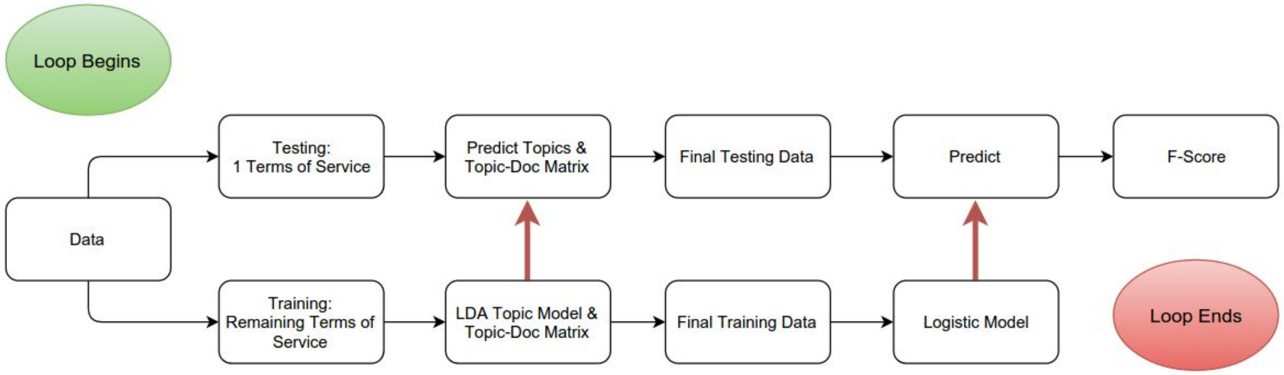
Since the size of our dataset was limited by the labor cost of manual annotation, we assumed that any measure of test error on a single held out validation set would have high variance. Thus, to get a more representative estimate of held-out test error, we performed a blocked k-fold cross-validation on our predictive model, using each company's ToS as a held-out test fold. This approach had the advantage of allowing us to easily determine the effect of inter-rater agreement on test accuracy, since each company has a unique value of kappa according to the two raters who classified it. We then performed the model fitting and validation process for each test fold and stored the test error metrics for later analysis.

For each validation, the data was divided by selecting one company's ToS to withhold as the test set and the remaining ToS's as the training set. The paragraphs in the training set were used to develop the LDA model, and this model was used to predict document-topic proportions for the test set of paragraphs. This ensured no leakage of information from the test set to the training set through the LDA. The topic proportions for each paragraph were then added as regressors to both the training and test sets.

Using the engineered features added in the preprocessing stage and the topic proportions from the LDA model, a logistic regression model was fit to the training set. Another logistic regression was fit without using the LDA proportions, in order to determine their utility in prediction. Since we were adding 50 variables with LDA and our training sizes were only in the hundreds, we wanted to be sure we weren't overfitting by including them. These logistic models were then used to predict a continuous response for the test set, and paragraphs with $p > 0.3$ were marked positive, using the conservative threshold we adopted in the exploration phase. Evaluation metrics of accuracy, recall, and precision, and F-score were then calculated from these predictions.

We also performed this same process for the subset of ToS that each rater annotated using the rater's individual ratings as the response. This gave us a 6-fold cross validation on each subset and provided an avenue to analyze the performance of the model in the absence of rater variability. If the model performed better even on these smaller training sets, we would have good evidence that the model would benefit from more consistent annotations.

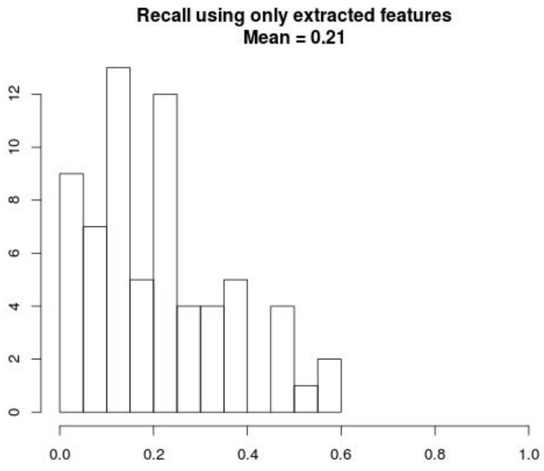
Below is a schematic of the predictive model, encapsulated in a loop representing the k-fold process utilized in both validations:

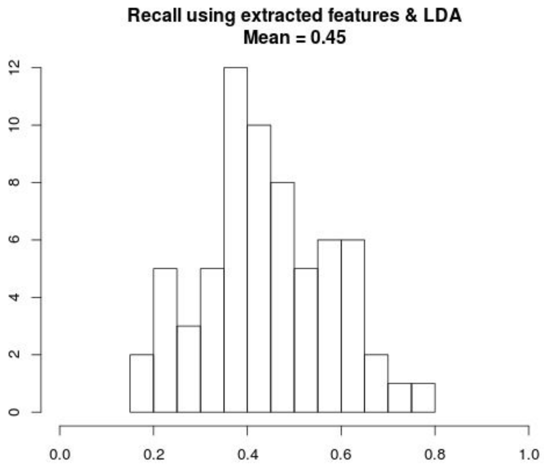


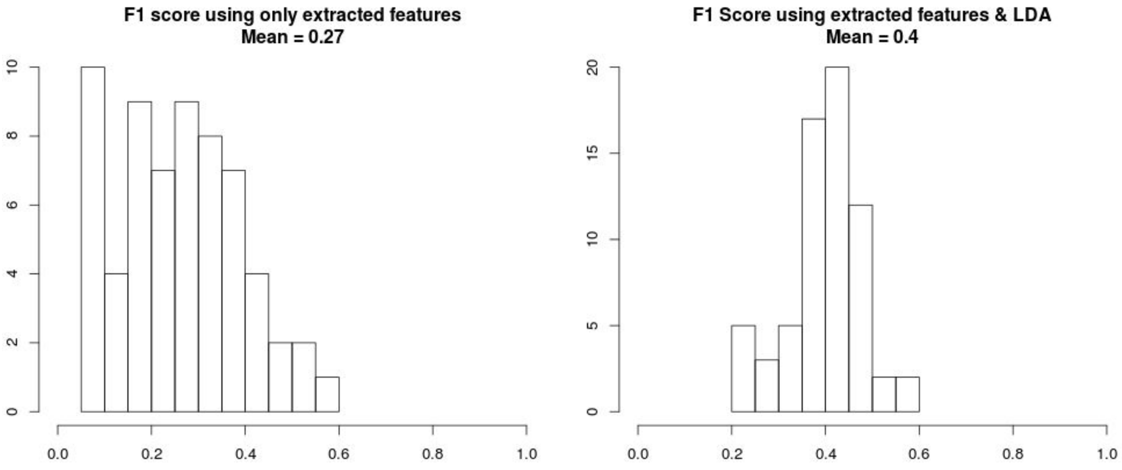
Evaluation

- **How well does your approach perform according to the metrics you describe in the Objectives section?**

Upon performing the validation for each company, it was found that the model using the LDA topics performed significantly better according to recall and F-score. Thus, we employed logistic regression using the engineered features and the LDA topic proportions as our final model. Below is a histogram comparing the models using each of the 66 pairs of companies as test sets in the same validation procedure as above to get a clearer idea of the distribution of test errors. The model including LDA is the clear winner.







Below is a table of our accuracy metrics from the single-company k-fold validation, ranked by F-score:

Company	Accuracy	Precision	Recall	F
Instagram	0.75	0.57	1.00	0.73
Cloudant	0.85	0.40	0.80	0.53
GitHub	0.76	0.58	0.47	0.52
Netflix	0.77	0.35	0.67	0.46
Yahoo	0.82	0.35	0.67	0.46
iCloud	0.78	0.36	0.55	0.44
Youtube	0.92	0.40	0.40	0.40
Google	0.78	0.75	0.27	0.40
SoundCloud	0.83	0.35	0.40	0.37
Twitter	0.78	0.38	0.33	0.35
Amazon	0.89	0.67	0.22	0.33
Facebook	0.85	0.22	0.25	0.24
Average	**0.82**	**0.45**	**0.50**	**0.44**

The results show high variance, likely due to our small test sample size, but the model methodology shows promise due to the favorable F-score values.

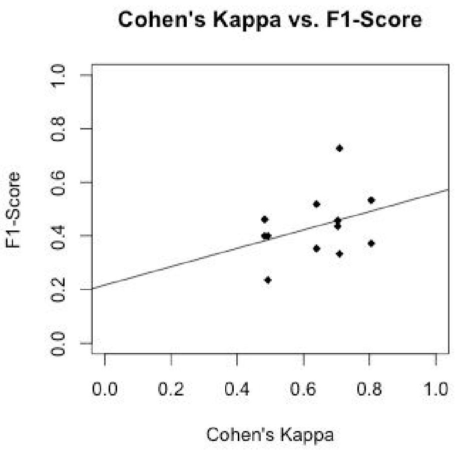
Below are the results of the single-company k-fold validation performed on each user's subset of the annotations:

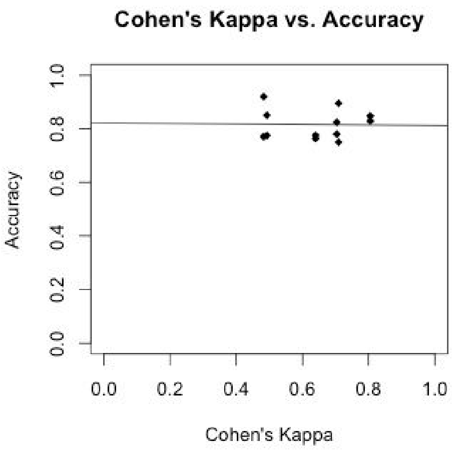
Rater	Accuracy	Precision	Recall	F
HM Average	0.70	0.45	0.65	0.51
KS Average	0.71	0.43	0.55	0.48
MH Average	0.79	0.34	0.48	0.46
JL Average	0.67	0.32	0.49	0.36
Average	**0.72**	**0.39**	**0.55**	**0.45**

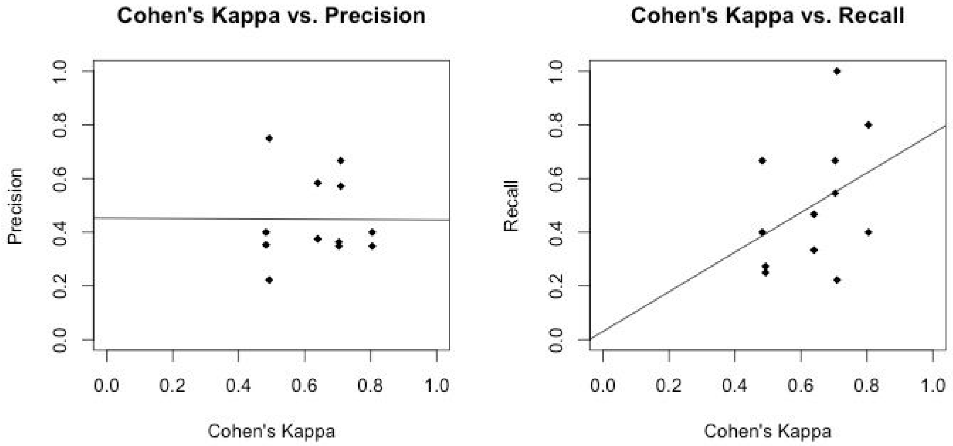
On average, recall is slightly higher, while the F-score is unaffected, despite training sets being roughly less than half the size of the prior validation on the whole data set (5 ToS's vs. 11). This indicates possible promise for improvement through greater annotation consistency, since each subset was reviewed by one person.

- **In what situations does your approach perform well? Where does it break down?**

Measuring by our most important metrics, recall and F-score, our methodology was most effective when Cohen's kappa was highest between the two raters. In the plots below, the F-score and recall seem to have a linear relationship with kappa. For example, the Cohen's kappa term between the raters for Cloudant and Instagram were 0.8 and 0.71 respectively, which corresponded to the two highest F-scores. Inversely, F-score and recall were lower when Cohen's kappa was lower. For example, a Cohen's kappa value of 0.49 corresponded to an F-score of 0.24 for Facebook. These observations further indicate that our model would be strengthened by expert annotation.







- **How does your approach stack up against other known approaches?**

Considering that applying machine learning techniques to ToS is novel, we were unable to directly compare our results to previous approaches. The closest methodology, the TOSDR, was difficult to directly compare against because the highlighted ToS components called out on the site were not mapped to ToS paragraphs.

Conclusions and Recommendations

- **What have you learned by doing this work?**

We learned that having consistent ground truth is critical to the success of a predictive algorithm. Although this is generally not a problem given the nature of ground truth, our actual values were based on a partially subjective annotation of the documents. Since we do not have legal expertise, the paragraphs each person labeled as important may have exhibited inconsistencies. As a result, noise may have been introduced to the predicted responses.

- **What are your final recommendations with regard to addressing the problem you have identified?**

To improve our model, we recommend employing two legal experts to manually label our terms of service agreements. We then recommend calculation of Cohen's kappa and iteratively completing the labelling process until a satisfactory agreement measurement is reached. We also recommend employing a third legal expert to decide a category for particular paragraphs where the two experts disagree. We believe this approach would mitigate variability.

Additionally, we recommend encapsulating this model in an application that is either employed on a company's ToS webpage or as a stand-alone service where a user could copy and paste any ToS. To explore possible designs, we created a Shiny App which allows the user to choose their threshold value and read the paragraphs that were indicated as important above the selected threshold. This application would be able to help the individual gain insight into the document, efficiently parse the document, and focus on particular sections. Any utilization of this application would be a marked improvement from the current environment where few users read, let alone understand, the ToS they agree to.