Using Sentiment Information for Preemptive Detection of Harmful Comments in Online Conversations

Éloi Brassard-Gourdeau[†], Richard Khoury[‡,*]
[†] Two Hat Security Research Corp., British-Columbia, Canada
[‡] Departement of Computer Science and Software Engineering, Université Laval, Québec, Canada

Abstract

The challenge of automatic moderation of harmful comments online has been the subject of a lot of research recently, but the focus has been mostly on detecting it in individual messages after they have been posted. Some authors have tried to predict if a conversation will derail into harmfulness using the features of the first few messages [1]. In this paper, we combine that approach with previous work on harmful message detection using sentiment information [2], and show how the sentiments expressed in the first messages of a conversation can help predict upcoming harmful messages. Our results show that adding sentiment features does help improve the accuracy of harmful message prediction, and allow us to make important observations on the general task of preemptive harmfulness detection.

1. Introduction

Billions of messages are sent online every day and hidden among them are millions of harmful messages. For instance, studies have shown that 17% of internet users receive cyber-bullying messages (a type of harmful messages), with a disproportionate number of targets being women (19%), low-income people (24%), and homosexuals (34%) [3], and that 10% of people develop depressive or suicidal thoughts as a result of these messages [4]. This makes the development of accurate and efficient content moderation systems a high priority. Most of the studies so far have focused on single-line detection. In other words, the models developed look at each message individually and decide whether it should be classified as harmful or not. While such systems can be good at detecting harmful messages once they are written [5–7], they cannot predict whether upcoming messages in a conversation will feature harmful content or not. This ability to flag interactions for moderation before they turn harmful would be hugely beneficial both for community moderators, allowing them to intervene more quickly and efficiently, and for users, preventing them from being targeted by harmful messages in the first place. This is the goal of the task of preemptive detection. It is however impossible to do when considering only a single message in isolation, and requires a model of the entire conversation. While many attributes can be derived from an entire conversation to aid in this prediction task [1], we feel that one that is not being properly studied is the use of sentiment information. Our intuition is that the sentiments expressed early in a conversation can help predict more accurately if it will degrade into harmfulness later on or not. If true, this would run counter to previous findings in the area. For instance, the experiments of [1] found sentiment information to have no predictive power in this task.

The rest of this paper is structured as follows. After a review of the relevant literature in Section 2, we will quickly go over the sentiment detection tool in Section 3. In the same section we will also study how our sentiment features can be added to the system of [1] and present the resulting preemptive detection tool. We will conduct an in-depth analysis of our results when using the dataset of [1] in Section 4. To expand on this study, we then perform a second set of experiments on another dataset in Section 5. Finally, we draw conclusions on preemptive detection and the use of sentiment information in Section 6.

*Richard.Khoury@ift.ulaval.ca

2. Related Work

The challenge of harmful message detection in online conversations has been studied since 2012. Various topics have been covered, such as hate speech detection [8, 9] and cyberbullying detection [5, 10, 11], and many architectures have been adopted and trained successfully for this task, including SVMs [7, 8], logistic regressions [9], and neural networks [11]. However, even the most recent work only focuses on single-line detection, meaning determining whether a comment that has already been posted is harmful or not by itself and outside the context of the conversation where it appears.

In addition, a message's sentiment information has rarely been considered among the attributes for harmfulness detection, much less studied in depth. A subjectivity lexicon was used in the study of [7], but their conclusion was that something more than sentiment keywords would be needed to inform a system. The systems of [5] and [12] both included sentiment among their inputs, but the papers do not discuss the importance or impact of that information. On the other hand, in our previous study [2] we looked at using sentiment in the particular context of subversive harmful message detection, and found that including it gives a 3% improvement.

One of the first and only studies on harmful message prediction at the conversation level is that of [1]. The authors showed that certain features in the first messages of a conversation, such as the use of first or second person pronouns and the presence of certain politeness strategies, can help predict if that conversation will remain healthy or if it will degrade and lead to harmful messages later on. Their work inspired the authors of [13], who trained and tested an SVM using TFIDF-weighted unigrams and bigrams as well as a BiLSTM using their own word embeddings. They were ultimately dissatisfied with their results, however the fact they focused only on words and did not use more sophisticated features such as those in [1] may be the cause. Finally, the authors of [14] did hostility presence and intensity prediction on Instagram comment threads using a variety of features, ranging from n-grams and word vectors to user activity and lexicons. The features are used to train a logistic regression model with L2 regularization. The authors conclude that there are four main predictors for hostility: the post author's history of receiving hostile comments, the presence of user-directed profanity in the thread, the number of distinct users posting comments in that thread, and the amount of hostility so far in a conversation.

However, much as for harmful message detection, sentiment information has not received much attention in harmful message prediction. Of the three studies mentioned, only [1] discusses it, and they dismiss it as equivalent to random chance. However, we believe that the issue comes not from sentiment information, but from how they incorporated it into their system. The focus of this paper is to study in depth the use of sentiment information for the task of harmfulness prediction, and to determine the conditions in which it can be included and benefit the system.

3. Conversation Model

3.1. Sentiment Detection Tool

In our previous work [2], we implemented a sentiment detection system in order to study whether sentiment information can help detect harmful content in a subversive setting (where users deliberately misspell harmful words to mask them from keyword filters). We found that sentiment information did correlate with harmfulness, and could be used to improve the accuracy of harmful message detection systems, both in a normal and in a subversive setting.

The sentiment detection tool implemented in that paper, which we will reuse in this one, is heavily inspired by previous works such as [15–17], where the authors used sentiment lexicons, such as SentiWordNet or General Inquirer, to detect the sentiment of a message. Our tool combines three popular lexicons, namely SentiWordNet[1], Afinn[2] and Bing Liu[3]. The authors found previously that these three lexicons have different strengths and weaknesses, and thus complement each other well. SentiWordNet is the biggest lexicon and assigns a positive and negative score between 0 and 1 to each word. Afinn assigns a single score between -5 and 5, with scores under zero meaning the words are negative. The Bing Liu lexicon has a positive and a negative word list. The lexicons are combined by splitting each into lists of positive and negative words for each of four parts-of-speech (noun, verb, adverb, and adjective), and normalizing the sentiment scores between 0 and 1.

Our sentiment detection tool begins by detecting sentiment-carrying idioms in the messages. For example, while the words "give" and "up" can both be neutral or positive, the idiom "give up" has a clear negative sentiment. Several of these idioms can be found in our lexicons, especially SentiWordNet (slightly over $60,000$). When detected, these idioms are marked so that our algorithm will handle them as single words. Next, it uses the NLTK *wordpunkt_tokenizer* to split messages into words, and the *pos_tagger* to get the part-of-speech of each word. Each word is then assigned a positive and a negative score, which is the sum of the score it has in the positive and negative lists of each of the three lexicons. A message is represented by the score of its three most positive words and its three most negative words. This gives us a total of 6 sentiment features for each message. For more details as to why the tool is built this way, please refer to [2].

3.2. **Model and features**

The authors of [1] split their conversation pragmatic features into two categories: 13 politeness strategies and 6 rhetorical prompts. The first category focuses on the use of politeness, such as greetings, gratitude, or the use of "please", and of impoliteness, such as direct and strong disagreement or personal attacks. The second category captures six domain-specific conversation prompts, which are six clusters of conversations discovered by an unsupervised technique trained on a different dataset that includes similar types of discussions. A new message's distance to each of these six clusters gives the six prompt features. This gives a total of 19 features per message, and the authors compute them for the first two messages of a given conversation, thus getting a set of 38 features. Using these features, the authors train a logistic regression model to predict if a conversation will derail into harmfulness based on its first two messages. The authors have made their code available publicly[4]. More details on these features and the regression model built from them can be found in the original article.

Our version of the model builds upon theirs by adding the 6 sentiment features measured by our sentiment tool for the same two messages. We also computed another sentiment feature representing the overall tone of the first two messages. This feature is computed by taking the sum of positive word scores of the first message and subtracting the sum of negative word scores to determine if the message is overall positive or negative, doing the same for the second message, and determining if the conversation starts with two positive messages, a positive followed by a negative, a negative followed by a positive, or two negative messages. This information is encoded as a one-hot vector[5] of length 4. In total, there

[1] http://sentiwordnet.isti.cnr.it/
[2] https://github.com/fnielsen/afinn
[3] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[4] https://github.com/CornellNLP/Cornell-Conversational-Analysis-Toolkit
[5] A vector composed entirely of 0s save for one "hot" element marked as 1.

are thus 38 text features from [1] and 16 sentiment features we added, for a total of 54 conversation features.

3.3. **Data and Training**

The dataset created for [1], which is available publicly along with their code, is a set of user conversations taken from the edit pages of English Wikipedia. The authors used Persepctive API[6] to pre-filter the conversations and keep only the ones with potentially harmful content. They further filtered to keep those conversations that started in a civil way, meaning that they did not have any harmful content in the first two messages. Moreover, they required conversation pairs, one derailing and one staying civil, from each Wikipedia page. This resulted in 1,270 conversation pairs from 582 different pages with an average length of 4.6 messages.

Our model is trained using *Scikit-learn*'s *LogisticRegression* and *SelectPercentile*, with a grid search on hyperparameters C between 10^{-4} and 10^4 and *percentile* between 10 and 100[7]. Training was done using a 5-fold cross validation. Apart from increasing the number of folds from 3 to 5 for more consistency between runs, all the training parameters are exactly the same as the ones in [1].

4. **Results and Analysis**

As in [1], our experiments consist in taking a pair of conversations, looking at their first two messages, and predicting which of the two conversations will remain healthy and which one will derail into harmfulness. All the results presented in the following section are the average of 10 separate runs, where we randomized the data split.

4.1. **Sentiment Features**

Our first experiment considers the predictive accuracy of sentiment information alone. In fact, the authors of [1] did include the sentiment lexicon of [18] in their research, and used it to extract two sentiment features per message. Their features were “has negative" and “has positive", each being 1 if a negative or positive word from the lexicon was present in the message and 0 otherwise. However, after testing these features, they concluded that sentiment was barely better than random chance at predicting harmful messages, and they did not include them in their set of 38 pragmatic features.

The goal of our first experiment is thus to validate that the sentiment features are in fact predictors of upcoming harmful messages. We trained and tested the model using four setups: using the original sentiment features of [1], our sentiment word features, our tone features, and all sentiment features combined. The results are presented in Table 1.

Table 1. Prediction accuracy using sentiment features.

Test	Features	Accuracy
Original sentiment	4	51.3
Our sentiment	12	55.7
Our tone	8	50.8
All features	24	55.8

Our results firstly confirm that the minimalist sentiment features of [1] are nearly equivalent to a random chance guess. This is likely due to the fact that over 70% of the messages

[6] https://www.perspectiveapi.com/

[7] C representing the regularization and *percentile* representing the percent of features to use.

containing a negative word also have a positive word, making it nearly impossible to discern a harmful message from a healthy one based on that information alone. Likewise, our tone information carries nearly no useful information. However, our more detailed word features do show an interesting predictive ability. Finally, combining all features together gives no gain compared to just using the word features; an unsurprising result, given that the other features seem to contain no predictive information.

This shows that, when it comes to sentiment information, it is not the overall sentiment of a message that is useful, but individual words. That level of detail is missing from both the original sentiment features (which only indicated whether positive or negative sentiment exist) and our tone features (which only indicate whether positive or negative sentiment is stronger). It is however present in the sentiment word features, which indicates the sentiment of the three most positive and most negative words of each message without making a judgment on whether the message overall is positive or negative. That finer level of granularity seems to be where the predictive information is found.

From this point forward, we will drop the sentiment of [1] and our tone features from our model, since they do not seem to be predictive of harmfulness. This will leave 12 sentiment features and a total of 50 conversation features.

4.2. **All Features**

Our next experiment consists in training and testing our model with and without the sentiment features. The goal is to highlight the gain in prediction accuracy that comes from including sentiment features. The results of that experiment are given in Table 2.

Table 2. Prediction accuracy with and without sentiment features.

Test	Features	Accuracy
Text features	38	58.6
Text + sentiment	50	60.5

In all 10 runs of our test, we found that the model including sentiment features consistently performs better than the one without. Our results using text features alone are consistent with those of [1], and adding sentiment features improves the prediction on average by 2%. This is consistent with the findings in [2], where it was found that sentiment information improved harmfulness prediction by 3%.

4.3. **Predictive Features**

It is interesting to examine which sentiment features contribute the most information to the prediction of how a conversation will develop. To do this, we take the average norm of the coefficient score of the logistic regression for each of the 50 features over the 10 runs of our experiment. The most informative features are simply those with the highest positive or negative coefficients, while features with coefficients around 0 have no influence on the prediction.

We found that the most predictive features were consistent from run to run. They are listed in Figure 1, along with their average coefficients. The top pragmatic features found match those identified in [1]. In addition to those, four of the sentiment features are among the 14 most predictive features found by the regression model.

For predicting conversations that will feature harmful messages, the strength of the first and second most negative words in the first message and of the third most negative word in the second message are all strong predictors. This indicates that strong negative words in both first messages will likely cause the conversation to degrade. Combined with the fact

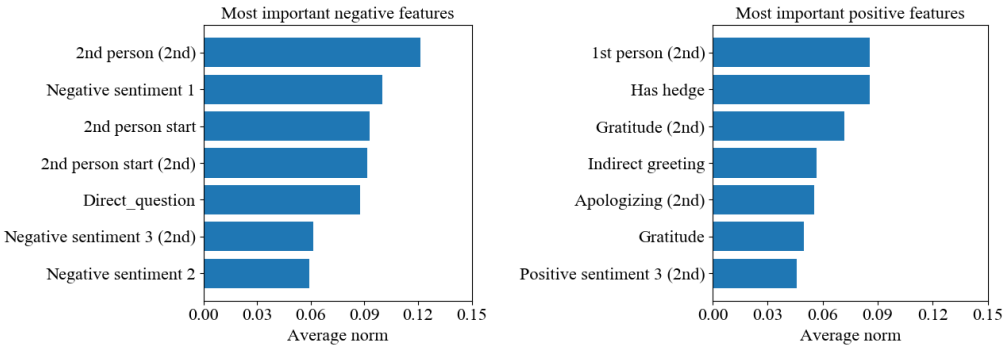

Figure 1. Feature importance when using 3 positive sentiment features and 3 negative sentiment features. The "(2nd)" refers to the feature on the second message, while its omission refers to the first message.

that second-person pronoun use in both messages are also strong harmfulness predictors, this may indicate conversations that begin with directed negative sentiments towards other participants.

On the other hand, only one of the sentiment features is among the strongest predictors of whether a conversation will remain healthy. It is the strength of the third most positive word in the second message. This is an interesting difference with the harmful case: while strong negative words are clear predictors of upcoming harmful messages, strong positive words are not predictors of healthy messages, but lower-ranked positive words are. This may indicate that abundance, not strength, of positive sentiment is what matters to predict health.

In order to verify that hypothesis, we re-trained and re-tested our model several times using between 1 and 7 positive or negative sentiment features. The best combination we found is using 5 positive sentiment features and only 2 negative ones, and this 56-feature model offers an average improvement of 1% on prediction accuracy compared to the 50-feature model of Table 2. The most predictive features in that test are shown in Figure 2. For harmfulness prediction, nothing has changed, save for the fact the third negative word of the second message has disappeared (as the feature is no longer part of the model) and the second negative word of the second message becomes the seventh most predictive feature (it was eighth previously). For health prediction, we can see that the newly-added features of the fourth and fifth positive words of the first message are now among the top predictors, beating out the third positive word from Figure 1. This confirms our earlier intuition.

4.4. **Case Studies**

Figure 3 has an example of the first two messages of a conversation that was mispredicted as healthy using the pragmatic features alone, but was correctly predicted as leading to harmful messages by our classifier with sentiment features.

The first message uses the first person and apologizes, both text features that predict a healthy conversation, and no other predictive text features are present in either message. As a result, the text-based system predicts they will lead to a healthy conversation. In reality, this conversation eventually degrades into the users attacking each other with messages such as "[username] actually blames others", "it's your problem", "you are just trying to find an excuse to take jabs at me" and eventually "[username] shut up".

When taking sentiment information into account, the picture is quite different. Both messages contain only a single strong positive sentiment word, the word "pretty" (score of

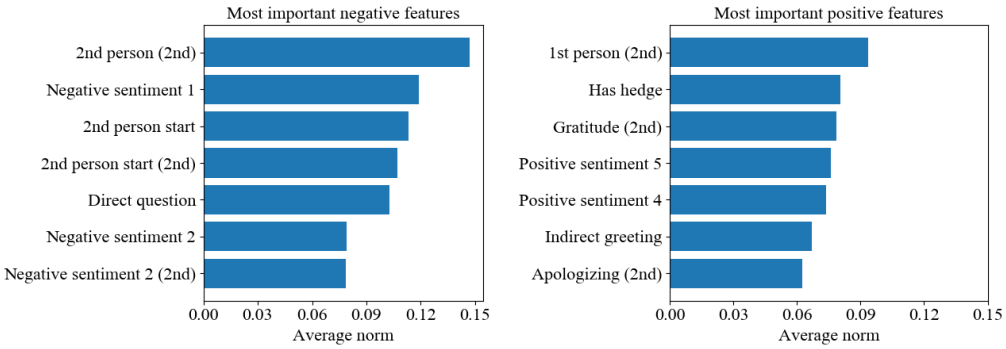

Figure 2. Feature importance when using 5 positive sentiment features and 2 negative sentiment features.

(1) I'm sorry to say it, but I'm pretty sure this is the only option left. This discussion has been so repetitive it's unbelievable. The mediation cabal has all but ceased, and the mediation of this talk page has **failed**. The RFC also did not work. I can see no other way to reslve the issue other than ArbCom. What does everyone else think?
(2) It's a pretty **useless** process. Mostly Admins listing Hebrew as a language, or displaying Israeli symbols on their user pages will respond they have no **problem** with the biased edits. **Worst**-case they ban you for suggesting the article needed comment or some type of oversight.

Figure 3. First two messages of a derailing conversation, with major good words underlined and major bad words in bold.

0.59). The other positive words are very weak, and the fourth and fifth positive words of the first message are "all" and "think" (scores of 0.04 and 0.02 respectively). On the other hand, the first message has two strong negative words, "failed" (score of 0.47) and "unbelievable" (score of 0.46), and the second message has three even stronger ones, "useless", "problem" and "worst" (scores of 0.67, 0.60, and 0.78 respectively). Negative features dominate in these messages, and as a result our model predicts correctly that this conversation will derail into harmfulness.

This example highlights one reason why the top positive words are not predictors of health: they can be used as modifiers to enhance negative words, as is the case of the word "pretty" in "pretty useless". We believe another reason the strongest positive words are not good predictors is sarcasm, which uses one or two very strongly positive words to convey a negative message. However, we found no examples of sarcasm in our dataset, so we could not confirm that hypothesis.

(1) not vandilism
(2) well sorry about replacing bands.but you **dumb cunt** fireworks is also a punk pop band

Figure 4. First two messages of a derailing conversation, with major good words underlined and major bad words in bold.

The sample conversation of Figure 4 is an example of the impact of strong negative words. The second message in particular contains an apology (positive pragmatic feature),

the strong positive word "well" (score of 0.46), and uses the second person (negative pragmatic feature). However, most people will pinpoint the two negative words as the strongest indicators this conversation will degrade. In fact, if those words were removed from the message, it would become a much more civil conversation. This illustrates how one or two strongly negative words can change the tone of a message and the flow of a conversation.

5. **Gaming Chat Moderation**

To validate the generality of our results, we decided to apply our model to a completely different setting from Wikipedia talk pages: live in-game chat conversations from a popular video game[8]. This dataset consists of 26,964 different conversations of up to 50 messages, with most messages being very short, around 4 words only. This makes it very different from the Wikipedia dataset, in which conversations are on average less than 5 messages long but messages are on average 58 words long. The last message of each conversation was reported by a user, and then a decision was made by a community moderator to either take action on the reported message or ignore the report. The dataset is balanced, with 54% of messages moderated and 46% ignored.

There are several other significant differences with the Wikipedia dataset. Unlike an edit discussion which has a well-identified initial message, a gaming chat conversation begins when the chat room is created and is continuously ongoing after that, with players joining and leaving at will. The dataset's 50-message conversations are actually composed of the reported message and the previous 49 messages. Moreover, the Wikipedia dataset contains mostly two- to four-person conversations, while very often over a dozen players can chat simultaneously (together or in intertwined separate discussions) and be present in the 50-message conversation.

The purpose of this experiment is slightly different from the previous one: while we still want to determine if it is possible to predict if a conversation will derail into harmfulness (meaning in this case that it will need moderation) from earlier messages, and to measure which text and sentiment features are the strongest predictors of this, we are no longer working with conversation pairs. Consequently, instead of choosing which of two conversations is most likely to go awry, we predict for each conversation individually if it will go awry or not, which is a much harder problem. Moreover, since the first message in a conversation is not the first message of the chatroom, we are not making a prediction from the beginning of a conversation but from an arbitrary point in the middle of it. Finally, taking only the first two messages as before would represent on average 8 words, which is not enough information to make a prediction from. Consequently, we use instead the 10 messages prior to the reported comment to predict whether the unseen final message will be harmful and require moderator action or not. This is thus a true preemptive moderation challenge: based on 10 messages, we are predicting whether an unseen 11th message will be moderated or not.

We will use the same 19 pragmatic features and 7 (5 positive and 2 negative) sentiment features per message as before. However, with 10 messages instead of 2, this means our model will have 260 features as input instead of 50. Moreover, we expect that message chronology will be a lot more important in a 10-message sequence than with 2 messages. Consequently, we decided to try two different models. The first one is the same logistic regression model as before. The second model is a recurrent neural network, specifically a uni-directional GRU with a kernel of 40 and a linear layer taking the final state of the GRU and producing a binary output. A recurrent neural network is a natural choice for a

[8]The dataset was provided by Two Hat Research Corp. with permission from the gaming company. The data was pseudonymized and users have agreed to have their chat used for moderation purposes. The data can not be shared publicly due to its sensitive nature.

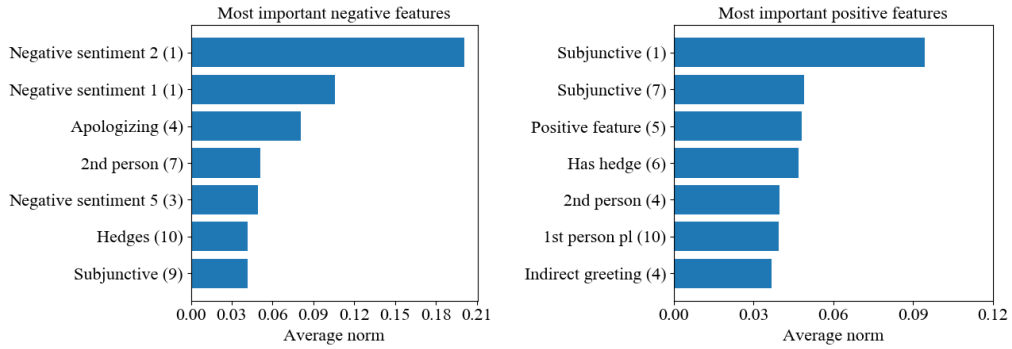

Figure 5. Feature importance in the gaming chat dataset. The number in parenthesis refers to the message's position before the reported message.

problem with a lot of features where chronology is important, and a similar model was used successfully in [2] for single-line harmful message detection.

The data was randomly split 70/20/10 into training/validation/test sets. We once again did 10 training and testing runs, using a different random split each time and 5-fold cross-validation within each run. Average results over all 10 runs are presented in Table 3. These results confirm that adding sentiment information helps improve the prediction of harmful conversations. The gain is greater for the logistic regression model, which in fact fails to make a prediction better than random chance without sentiment information. The RNN fares better, probably because it can better handle the large number and sequential nature of the features, but it still gains 1% by including sentiment information.

Table 3. Results for both models using text features alone (190 features) or text and sentiment features (260 features).

Model	Features	Accuracy	F1 score
Regression	190	50.9%	0.556
Regression	260	57.4%	0.564
RNN	190	60.6%	0.686
RNN	260	61.7%	0.691

As before, we use the average coefficient score of each feature over the 10 runs to rank the features by predictive importance. The top features are shown in Figure 5. There are some differences with the results of the Wikipedia test. Most notably, the "subjunctive" feature[9] is a strong predictor of healthy conversations in this experiment. Looking more closely, this feature is predictive of unmoderated conversations two-thirds of the times it appears; however, it appears in less than 1% of chat conversations. This difference is therefore not significant in practice.

On the other hand, the coherent aspects with the previous experiment are very interesting. In both experiments, the features "has hedge"[10], the use of 1st person pronouns, and indirect greetings[11], are indicators of healthy conversations, while strong negative-sentiment words are indicators of an upcoming harmful comment that will need to be moderated. Moreover, unlike with the "subjunctive" feature, these features all occur in a significant number of the

[9]Expressions such as "would you" and "could you".
[10]"Has hedge" refers to the presence of hedges, or mitigating words, like "think", "almost", "rather", etc. This differs from the feature "hedges", which looks for dependencies and requires the subject of the message to express this hedge.
[11]The presence of words like "hey", "hello" or "hi".

conversation. This indicates that the method is generalizable and can be applied to different types of online conversations.

Next, we considered the question of which messages in the conversation contain the most predictive features. To this end, we considered the 26 (10%) most predictive positive and negative features, and grouped them per message. The results, given in Table 4, show that features predicting both health and harmfulness can be found throughout the conversation. However, while health predictors are distributed evenly in the conversation, harmfulness predictors are concentrated in the final three messages. This indicates that a healthy conversation is an ongoing process, but a few bad messages can very quickly turn the tides of the conversation and lead to harmful messages being posted. This also indicates a limit to preemptive moderation using this method: long-term predictions are not valid, and one must focus on clues in the latest messages. To confirm this, we ran the experiment again using only 3 messages before the reported message instead of 10. The results are almost identical to before: the logistic regression classification has an accuracy of 57.7% with sentiment features and 51.7% without, while the RNN has an accuracy of 61.8% with sentiment features and 60.9% without. It seems clear, then, that the previous seven messages did not contribute significantly to the prediction accuracy.

Table 4. Number of positive and negative predictive features per message before the reported message.

Message	10	9	8	7	6	5	4	3	2	1
Positive	3	3	3	2	1	2	4	3	3	2
Negative	2	2	1	3	2	1	3	6	2	4

6. **Conclusion**

In this paper, we studied how sentiment information can be used as a feature for the task of predicting harmful messages in online conversations. We conducted this study using two very different online conversation datasets. The results of our experiments allow us to draw some important conclusions that can guide both future research and practical implementations of preemptive moderation tools:

(1) Sentiment information is indeed a predictor of harmful messages. Using it improves a system's performance by between 1% and 6%, which is consistent with previous results in [2]. This notably runs counter to previously-published results from other authors that indicated that sentiment information performs no better as a predictor of harmful messages than random chance.
(2) Sentiment information is found at a fine granularity, at the individual word level. Using coarser information, such as overall message sentiment, is not informative. This may explain the above-mentioned contrary previously-published results.
(3) It takes a lot of weak positive words to maintain a healthy conversation, but only a few strong negative words can turn a conversation harmful.
(4) The features that are predictive of health and harmfulness are consistent between two very different formats of conversations, and a preemptive detection system may therefore be generalizable to multiple different online communities.
(5) A conversation turns negative very quickly, and consequently negative predictors are concentrated in the most recent messages. This may put a natural limit to the range of preemptive detection. This range limit seems to be of 3 messages in our results.

The tasks of harmful message prediction is still in its infancy, and there is still a lot of room for research. For example, work so far has focused on using regular conversation

features as predictors. Future work could look at adding community-based features such as the reactions of community members to enrich the model by creating a more complete picture of online life.

Acknowledgments

This research was made possible by the financial, material, and technical support of Two Hat Security Research Corp., and the financial support of the Canadian research organization MITACS.

References

[1] J. Zhang, J. Chang, C. Danescu-Niculescu-Mizil, L. Dixon, Y. Hua, D. Taraborelli, and N. Thain. "Conversations Gone Awry: Detecting Early Signs of Conversational Failure". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1350–1361. DOI: 10.18653/v1/P18-1125. URL: https://www.aclweb.org/anthology/P18-1125.

[2] E. Brassard-Gourdeau and R. Khoury. "Subversive Toxicity Detection using Sentiment Information". In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 1–10. URL: https://www.aclweb.org/anthology/W19-3501.

[3] D. W. Hango. *Cyberbullying and cyberstalking among Internet users aged 15 to 29 in Canada*. https://www150.statcan.gc.ca/n1/pub/75-006-x/2016001/article/14693-eng.htm. Statistics Canada, 2016.

[4] A.-D. League. *Free to Play? Hate, Harassment, and Positive Social Experiences in Online Games*. https://www.adl.org/free-to-play. ADL Report, 2019.

[5] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali. "Mean Birds: Detecting Aggression and Bullying on Twitter". In: *CoRR* abs/1702.06877 (2017). arXiv: 1702.06877. URL: http://arxiv.org/abs/1702.06877.

[6] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos. "Deeper Attention to Abusive User Content Moderation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 1125–1135. DOI: 10.18653/v1/D17-1117. URL: https://www.aclweb.org/anthology/D17-1117.

[7] C. V. Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. D. Pauw, W. Daelemans, and V. Hoste. "Automatic Detection of Cyberbullying in Social Media Text". In: *CoRR* abs/1801.05617 (2018). arXiv: 1801.05617. URL: http://arxiv.org/abs/1801.05617.

[8] W. Warner and J. Hirschberg. "Detecting hate speech on the world wide web". In: *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics. 2012, pp. 19–26.

[9] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. "Abusive language detection in online user content". In: *Proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee. 2016, pp. 145–153.

[10] K. Reynolds, A. Kontostathis, and L. Edwards. "Using machine learning to detect cyberbullying". In: *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on*. Vol. 2. IEEE. 2011, pp. 241–244.

[11] S. Agrawal and A. Awekar. "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms". In: *CoRR* abs/1801.06482 (2018). arXiv: 1801.06482. URL: http://arxiv.org/abs/1801.06482.

[12] H. Dani, J. Li, and H. Liu. "Sentiment Informed Cyberbullying Detection in Social Media". In: *Machine Learning and Knowledge Discovery in Databases* (Jan. 2017), pp. 52–67.

[13] M. Karan and J. Šnajder. “Preemptive Toxic Language Detection in Wikipedia Comments Using Thread-Level Context”. In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 129–134. URL: https://www.aclweb.org/anthology/W19-3514.

[14] P. Liu, J. Guberman, L. Hemphill, and A. Culotta. “Forecasting the presence and intensity of hostility on instagram using unsing linguistic and social features”. In: *Twelfth International AAAI Conference on Web and Social Media*. 2018.

[15] B. Ohana, S. J. Delany, and B. Tierney. “A case-based approach to cross domain sentiment classification”. In: *International Conference on Case-Based Reasoning*. Springer. 2012, pp. 284–296.

[16] F. Å. Nielsen. “A new ANEW: Evaluation of a word list for sentiment analysis in microblogs”. In: *arXiv preprint arXiv:1103.2903* (2011).

[17] P. Tumsare, A. S. Sambare, S. R. Jain, and A. Olah. “Opinion mining in natural language processing using SentiWordNet and fuzzy”. In: *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume* 3 (2014), pp. 154–158.

[18] B. Liu, M. Hu, and J. Cheng. “Opinion Observer: Analyzing and Comparing Opinions on the Web”. In: *Proceedings of the 14th International Conference on World Wide Web*. WWW ’05. Chiba, Japan: ACM, 2005, pp. 342–351. ISBN: 1-59593-046-9. DOI: 10.1145/1060745.1060797. URL: http://doi.acm.org/10.1145/1060745.1060797.