Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 144–149
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Detecting Early Signs of Cyberbullying in Social Media

Niloofar Safi Samghabadi♠, A. Pastor López-Monroy♣, Thamar Solorio♠
♠Department of Computer Science, University of Houston, USA
♣Department of Computer Science, Mathematics Research Center (CIMAT), Mexico
nsafisamghabadi@uh.edu, pastor.lopez@cimat.mx, tsolorio@uh.edu

Abstract

Nowadays, the amount of users' activities on online social media is growing dramatically. These online environments provide excellent opportunities for communication and knowledge sharing. However, some people misuse them to harass and bully others online, a phenomenon called cyberbullying. Due to its harmful effects on people, especially youth, it is imperative to detect cyberbullying as early as possible before it causes irreparable damages to victims. Most of the relevant available resources are not explicitly designed to detect cyberbullying, but related content, such as hate speech and abusive language. In this paper, we propose a new approach to create a corpus suited for cyberbullying detection. We also investigate the possibility of designing a framework to monitor the streams of users' online messages and detects the signs of cyberbullying as early as possible.

Keywords: Cyberbullying Detection, Text Mining, Early Text Categorization

1. Introduction

In recent years, the internet has become the primary communication tool worldwide.[1] There are several social media platforms where people can share information and interact with each other in a virtually unlimited space. Although such platforms are beneficial for online users to develop their social skills and learn about new ideas and issues, they also put them under the risk of harassment, bullying, and cyber-attacks. Cyberbullying is defined as the use of information/communication technologies (ICT's) to harm others by sending or posting negative, harmful, false, or mean content to them *intentionally* and *repeatedly*. The most vulnerable groups targeted by this phenomenon are teens and pre-teens (Livingstone et al., 2010). Previous research shows that there is a statistically significant relationship between low self-esteem and experiences with cyberbullying (Patchin and Hinduja, 2010). Relevantly, cyberbullying victims have been reported to face various psychological and physical disorders that sometimes may lead them to harm themselves (Xu et al., 2012). Therefore, it is extremely important to detect cyberbullying incidents before they cause irreparable damages to the victims.

Several works have been done towards finding cyberbullying traces on social media by detecting online hateful and aggressive comments. Still, most of these efforts are focused on offline settings and only detect the event after it took place. Therefore, none of these methods can be used for prevention.

In this research, we aim to detect early signs of cyberbullying using *as few textual evidence as possible* by providing *timely* predictions. The main contributions of this work are listed as follows:

- A new methodology for creating a cyberbullying corpus and the first dataset suited for the task of early cyberbullying prediction.
- A new strategy to detect cyberbullying events as early as possible and the first evaluation framework that takes both the performance and the earliness of the predictions into account.

2. Related Research

Although there are several works on detecting different types of online aggression (Wulczyn et al., 2016; Nobata et al., 2016; Van Hee et al., 2018; Qian et al., 2018; Mishra et al., 2019a; Mishra et al., 2019b), only a few of them address cyberbullying detection. Dinakar et al. (2012) construct a common sense knowledge base - BullySpace - with knowledge about bullying situations and a wide range of common daily topics. Xu et al. (2012) study bullying traces and formulate cyberbullying detection as different Natural Language Processing (NLP) tasks. For instance, they use latent topic modeling to analyze the topics commonly discussed in bullying comments. Some previous works investigate cyberbullying on Instagram and Vine (Hosseinmardi et al., 2014; Hosseinmardi et al., 2015; Rafiq et al., 2018). For instance, Hosseinmardi et al. (2015) use a combination of textual, user-level, and image-related features to find cyberbullying incidents on Instagram media sessions. There are also a few studies that use time-related information to detect cyberbullying by using several different temporal features (Soni and Singh, 2018) and modeling the structure of a social media session with a hierarchical attention model (Cheng et al., 2019).

The main limitation of the previous systems is that they are built using an offline settings, and cannot detect cyberbullying in its early stages. Concerning this problem, early text categorization strategies could be a solution to model the dynamics of online conversations and provide timely predictions based on little evidence. Early text categorization is an emerging research topic which is being more popular, by reason of the specialized forums such as eRisk-CLEF.[2] eRisk started from 2017, and have emphasized topics such as detecting the early signs of depression (Losada et al.,

[1] http://www.gallup.com/poll/179288/new-era-communication-americans.aspx

[2] https://erisk.irlab.org

Q: didn't you used to make yourself throw up or something? It obviously didn't work because you're still over weight **A**: you're ignorant. **Q**: I'm not trying to be!!!! you're just better off dead so go right ahead. Nobody's holding you back honey. We won't miss you. **A**: thanks for the clarification **Q**: glad I could help! Let me know when you're dead so I can spit on your grave!!! :-) **A**: ok **Q**: Fucking bulimic bitch **A**: yeah totally!! **Q**: tell your mom I said hi when you see her in hell!!! She's so proud of how you've turned out. Just kidding **A**: she's definitely in heaven. and she's my god mother. and I know she loves me **Q**: oh look here your best friend coming to the rescue how cute. She secretly thinks you're worthless too. Nobody actually cares! They just say they do. Oh silly Meaghan so naive. You need serious help. Maybe you should ask your pointer and middle fingers? They've seemed to help you this far **A**: please just stop.

Table 1: Parts of a cyberbullying instance in our corpus.

2017a), anorexia (Losada et al., 2018; Losada et al., 2019), and self-harm (Losada et al., 2019) with monitoring the threads of online messages collected from Reddit.[3]

In this research, we investigate the possibility of employing the early text classification approach to tackle the problem of cyberbullying detection. We first introduce a new dataset suited for the task. Then, we conduct initial experiments to detect cyberbullying incidents as early as possible.

3. Data Collection

Abusive language detection can be considered as the initial step towards finding cyberbullying incidents. Cyberbullying happens when the victim receives several offensive messages repeatedly. Therefore, at least parts of the users' conversations should be monitored to detect such episodes. We collect our data from ask.fm.[4] This platform became the largest Q&A network in the world in 2017, reaching 215 million registered users. [5] ask.fm is a semi-anonymous social network that allows people to send comments/questions to any other user anonymously. This anonymity option provides the possibility for the attackers to freely harass users by sending lots of invective messages to their pages. Typically in ask.fm, the data consists of question-answer pairs in users' timeline.

Figure 1 shows the corpus creation scheme. We collect a large amount of ask.fm data, including the full history of question-answer pairs for 3K users. The question field includes a question/comment posted by the other users, and the answer field consists of the reply to that question/comment provided by the owner of the account. As we mentioned earlier, for finding the cyberbullying incidents, we may look for the threads of messages that include high ratio of abusive comments. We use our previous system for abuse detection on ask.fm (Samghabadi et al., 2017). We utilize the ask.fm corpus proposed in the same work for training the model and label each row of our data automatically. To make the cyberbullying instances, we create a fixed-length sliding window and move it through the whole history of question-answer pairs per user. For each window sample, we calculate the ratio of offensive questions/comments that the user received inside the window. If it is greater than a pre-defined threshold, we consider the window as a *potential* cyberbullying event. Additionally, we check whether we can expand the potential negative window by adding more question-answer pairs to it, yet keeping the inside negativity rate greater than the defined threshold. This step is crucial to capture the whole cyberbullying episode. Finally, since automatic labeling is likely to be noisy, we asked two annotators to manually check the resulting windows to assure that they represent real cyberbullying incidents. A window is tagged as cyberbullying, where both annotators agree that it includes a cyberbullying incident. Figure 1 shows some parts of a cyberbullying instance in our corpus. We empirically fixed the minimum window size and negativity threshold to 20 and 40%, respectively (i.e., the potential cyberbullying windows include at least 20 question-answer pairs from a specific user's timeline, and at least 40% of questions are labeled as offensive).

For the non-cyberbullying instances, we apply the same method, but inversely. In this case, we look for the windows that have the negativity ratio less than the defined threshold. We create bins of various negativity ratios (e.g., 0%-5%, 5%-10%, etc.) and make sure to add a fair number of samples from each category to our data. As for the false-positive examples, we also add the window samples that are labeled as highly negative but are not annotated as cyberbullying after manual checking (e.g., when two users send negative comments toward each other in the third user's timeline)

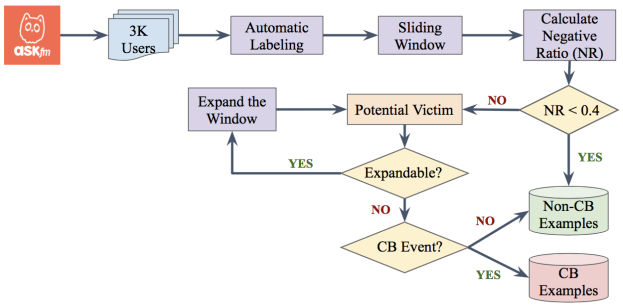


Figure 1: Overall process of building the new corpus.

Table 2 shows the distribution of the data in terms of the number of users in each class. Since *cyberbullying is a rare event*, we keep the ratio of positive to negative examples 1:10 to be closer to the real case scenarios. Finally, we divide all training and test examples to 10 different chunks to make the corpus suitable for early text classification. For every instance, each chunk contains 10% of all the question-answer pairs for that user.

[3] https://www.reddit.com

[4] https://ask.fm

[5] https://en.wikipedia.org/wiki/Ask.fm#2016\OT1\textendashpresent:_Purchased_by_Noosphere_and_new_cryptocurrency_plans

Class	training	test	Total
cyberbullying	19	8	27
non-cyberbullying	190	80	270
Total	209	88	297

Table 2: Statistics for our ask.fm data.

4. Methodology

Early text classification aims at developing a predictive model that is capable of determining the class that a document belongs to as early as possible, using partial information (Escalante et al., 2015). In this scenario, the instances (conversations) are read sequentially in chunks of texts that are fed into a classifier in an incremental fashion to obtain the prediction at chunk *t*. In our case, at every time *t*, we only have access to question-answer pairs in the first *t* chunks of test data to make the predictions. However, the training is done as usual (using all 10 chunks per instance). The intuition behind this scenario is to learn the overall pattern of a conversation and to investigate how helpful this pattern is to detect cyberbullying in the early stages of the conversation. This is the most standard framework for early prediction according to different forums such as eRisk (Losada et al., 2017a; Losada et al., 2019).

4.1. Feature Engineering

We use the following features to extract the information from the text:

Lexical: We use word n-grams (n = 1, 2, 3) and char n-grams (n = 3, 4, 5) as they are proven to be effective lexical representation for abuse and hate speech detection. For word n-gram features, we build a vocabulary that only considers top 10K features ordered by term frequency across the corpus. We weigh each term with its term frequency-inverse document frequency (TF-IDF).

Word Embeddings: The idea behind this approach is to map the words to a vector space model to improve lexical semantic modeling (Le and Mikolov, 2014). We use the pre-trained Google News word2vec model, including embeddings for about 3 million words. We create our feature vector by averaging the word embeddings of all the words in each post.

Style and Writing density (WR): This category extracts the stylistic properties of the text, and consists of the number of words, characters, all uppercase words, exclamations, question marks, as well as average word length, sentence length, and words per sentence.

LIWC (Linguistic Inquiry and Word Count): LIWC2007 (Pennebaker et al., 2007)) extracts different language dimensions like different emotions (e.g., sadness, anger, etc.), self-references, and casual words in each text. To create this feature set, we use a normalized count of words separated by any of the LIWC categories.

DeepMoji: The emojis are used to better understand the textual message by suggesting pictures that may help to represent it better. DeepMoji (Felbo et al., 2017) is a deep learning model that is pre-trained on a large set of Twitter data. Given an input text, this model provides an output representation for 64 frequently used online emojis. This representation shows how relevant each of those emojis is to the given input. We apply this pre-trained model on our data and extract the last hidden representation as the feature set for each post.

5. Experiments and Results

In the experiments, for each instance in our corpus, we have ten chunks, any of which includes 10% of question-answer pairs in that conversation. The first chunk contains the oldest 10% of the question-answer pairs, the second chunk consists of the second oldest 10%, and so forth.

5.1. Experimental Setup

In our chunk-by-chunk setting, we consider all questions and all answers within a chunk as the separate documents. Then, we extract the features from each document instead of a single post. The reason for separating questions and answers is that we believe these two categories of posts reflect two different views (i.e., commenters vs. account holder). We concatenate question-based and answer-based feature vectors to get a single representation for each instance. Then we feed these final representations to a linear SVM classifier. For each set of features, we tune the C parameter of the classifier with a grid search over values $\{0.1, 1, 2, 5, 10\}$.

5.2. Evaluation

For evaluating our early predictive model, we report the performance of the different methods using increasing amounts of textual evidence (chunk-by-chunk evaluation). More specifically, we evaluate the model in 10 consecutive iterations across the test set. In the first iteration, we generate a document representation starting with the first chunk, and then for each next iteration, we incrementally add one more chunk of data. The model makes predictions incrementally, as well. This chunk-by-chunk evaluation is a strategy that has been used to evaluate early classification models (Escalante et al., 2015; Errecalde et al., 2017; Losada et al., 2017b; Losada et al., 2018; López Monroy et al., 2018). As for the evaluation metric, we report F1-score for the cyberbullying class (the class of interest). We use this metric because the corpus is highly imbalanced towards the non-cyberbullying class.

5.3. Classification Results

Table 3 shows the classification results in terms of F1-score for the cyberbullying class. The results of WR and LIWC features are not included in the table due to the very low performance of the model using these features. Even combining these features with the other ones does not improve the performance. However, they seem to be helpful for the task of abusive language detection (Samghabadi et al., 2017). This contradiction indicates that in practice, there are some differences between the two tasks of abusive language and cyberbullying detection.

Based on the results, the best F1 measure is obtained from DeepMoji features using eight chunks of data. Even in earlier chunks, this method works significantly better than the other approaches. It shows that emoji-based representation for cyberbullying and non-cyberbullying instances are

Feature	ch1	ch2	ch3	ch4	ch5	ch6	ch7	ch8	ch9	ch10
Unigram	0.46	0.54	0.66	**0.76**	0.61	0.71	0.71	0.61	0.61	0.67
Bigram	0.00	0.20	0.00	0.20	0.36	0.36	**0.40**	0.40	0.22	0.40
Trigram	0.00	0.20	0.22	0.22	0.40	0.54	0.54	**0.61**	0.50	0.33
Char 3gram	0.40	0.22	0.40	0.40	**0.54**	0.36	0.40	0.54	0.54	0.54
Char 4gram	0.22	0.22	0.22	0.22	**0.40**	0.40	0.40	0.40	0.40	0.40
Char 5gram	0.00	0.22	0.22	0.22	0.22	**0.40**	0.40	0.40	0.40	0.40
Word2Vec	0.43	**0.59**	0.47	0.53	0.53	0.57	0.50	0.50	0.50	0.36
Unigram + Word2Vec	0.67	0.61	0.67	0.67	**0.71**	0.71	0.71	0.71	0.61	0.66
DeepMoji	0.73	0.78	0.75	0.80	0.88	0.82	0.82	**0.93**	0.75	0.77

Table 3: F1-score for the chunk-by-chunk evaluation for the positive class. The bold values show the best performance gained for each feature set.

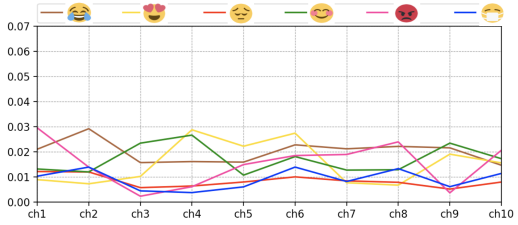
likely to be entirely different. We further analyze this result in Section 5.4..

Taking into account that the average number of question-answer pairs in each chunk of the test data is 4, unigram+Word2Vec and DeepMoji features show very promising results in the earlier chunks (considering only a few question-answer pairs). Overall, it seems that adding more information to the test data decreases the performance of the system after a while (especially in the last two chunks). Even for the Word2Vec feature, we get the best performance using only the first two chunks of the data. The reason could be the distribution of the offensive messages in a cyberbullying episode. These events are usually started with a couple of questions/comments from the attacker(s), and as they go forward, one or more users get involved in the conversation as the victim's bystanders. Some of these users try to encourage the victim to stay strong, and some others start defending the victim by posting aggressive comments targeting the attacker(s). This information possibly confuses the classifier when it gets access to the later chunks. To sum up, Table 3 shows that we can successfully adapt the early text categorization approach to the cyberbullying detection task, where the system shows better performance using less evidence.

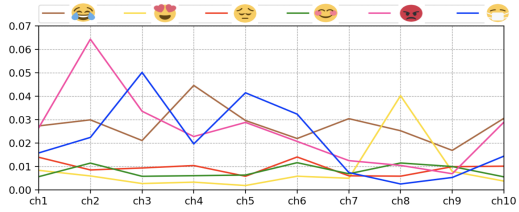
5.4. Analysis

Figure 2 illustrates the flow of emojis for a non-cyberbullying and a cyberbullying instance in our corpus. It helps us to understand better why DeepMoji representation helps detect early signs of cyberbullying. For making this figure, we choose 6 out of 64 emojis from the output of the DeepMoji model. We try to select an emoji set that covers various emotions (e.g., happiness, sadness, anger). Then, we plot the probability of each emoji to be related to the textual data we have available in each chunk.

Based on Figure 2a, in a non-cyberbullying thread, we have a mixture of the emojis (i.e., overall, no emoji is dominant). But in a cyberbullying one (Figure 2b), negative emojis like 😡 and 😷 are almost dominant, specifically in the first few chunks. It is interesting to see that laughing face (😂) is also showing a higher probability in this case. So, we can conclude that probably in this instance, the attacker(s) makes fun of the victim.



(a) Non-cyberbullying instance



(b) Cyberbullying instance

Figure 2: Flow of Emojis.

6. Conclusion

In this paper, we present a new approach to create a linguistic resource for detecting the early signs of cyberbullying. We start by automatically labeling all rows of data. Then, we move a sliding window through the history of each user's interactions to find the potential cyberbullying cases based on the ratio of received abusive messages. Finally, each of these possible cyberbullying instances is annotated manually to make sure that it includes a cyberbullying incident. We follow the same process to label the non-cyberbullying class. Furthermore, we use a simple set of lexical, semantic, and stylistic features to train an SVM classifier for cyberbullying detection. This system is evaluated over the different chunks of test data iteratively. The final results demonstrate that early text classification scenarios can be successfully adapted to detect cyberbullying at the early stages.

For future work, we plan to enrich our ask.fm corpus by collecting more users. Also, instead of chunk-by-chunk evaluation, we plan to examine the post-by-post evaluation that is closer to the real case scenario. Our ultimate goal is to design a sequential decision-making module, which can provide accurate and timely predictions on whether to label a conversation as cyberbullying based on the current information, or wait for more evidence.

7. Bibliographical References

Cheng, L., Guo, R., Silva, Y., Hall, D., and Liu, H. (2019). Hierarchical attention networks for cyberbullying detection on the instagram social network. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 235–243. SIAM.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

Errecalde, M. L., Villegas, M. P., Funez, D. G., Ucelay, M. J. G., and Cagnina, L. C. (2017). Temporal variation of terms as concept space for early risk prediction.

Escalante, H. J., Montes-y Gómez, M., Villaseñor-Pineda, L., and Errecalde, M. L. (2015). Early text classification: a naïve solution. *arXiv preprint arXiv:1509.06053*.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *2017 Conference on Empirical Methods in Natural Language Processing-Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hosseinmardi, H., Ghasemianlangroodi, A., Han, R., Lv, Q., and Mishra, S. (2014). Analyzing negative user behavior in a semi-anonymous social network. *CoRR abs*, 1404.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., and Mishra, S. (2015). Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196.

Livingstone, S., Haddon, L., Görzig, A., and Ólafsson, K. (2010). Risks and safety on the internet. *The Perspective of European Children. Final Findings from the EU Kids Online Survey of*, pages 9–16.

López Monroy, A. P., González, F. A., Montes, M., Escalante, H. J., and Solorio, T. (2018). Early text classification using multi-resolution concept representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT-2018, Volume 1 (Long Papers)*, pages 1216–1225. Association for Computational Linguistics.

Losada, D. E., Crestani, F., and Parapar, J. (2017a). Clef 2017 erisk overview: Early risk prediction on the internet: Experimental foundations. In *CLEF (Working Notes)*.

Losada, D. E., Crestani, F., and Parapar, J. (2017b). erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 346–360. Springer.

Losada, D. E., Crestani, F., and Parapar, J. (2018). Overview of erisk 2018: Early risk prediction on the internet (extended lab overview). In *Proceedings of the 9th International Conference of the CLEF Association, CLEF*.

Losada, D. E., Crestani, F., and Parapar, J. (2019). Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer.

Mishra, P., Del Tredici, M., Yannakoudakis, H., and Shutova, E. (2019a). Abusive language detection with graph convolutional networks. *arXiv preprint arXiv:1904.04073*.

Mishra, P., Del Tredici, M., Yannakoudakis, H., and Shutova, E. (2019b). Author profiling for hate speech detection. *arXiv preprint arXiv:1902.06734*.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Patchin, J. W. and Hinduja, S. (2010). Cyberbullying and self-esteem. *Journal of School Health*, 80(12):614–621.

Pennebaker, J. W., Booth, R. J., and Francis, M. E. (2007). Liwc2007: Linguistic inquiry and word count. *Austin, Texas: liwc.net*.

Qian, J., ElSherief, M., Belding, E., and Wang, W. Y. (2018). Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123, New Orleans, Louisiana, June. Association for Computational Linguistics.

Rafiq, R. I., Hosseinmardi, H., Han, R., Lv, Q., and Mishra, S. (2018). Scalable and timely detection of cyberbullying in online social networks. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 1738–1747. ACM.

Samghabadi, N. S., Maharjan, S., Sprague, A., Diaz-Sprague, R., and Solorio, T. (2017). Detecting nastiness in social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 63–72.

Soni, D. and Singh, V. (2018). Time reveals all wounds: Modeling temporal characteristics of cyberbullying. In *Twelfth International AAAI Conference on Web and Social Media*.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., and Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one*, 13(10).

Wulczyn, E., Thain, N., and Dixon, L. (2016). Ex machina: Personal attacks seen at scale. *CoRR*, abs/1610.08914.

Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12,

pages 656–666, Stroudsburg, PA, USA. Association for Computational Linguistics.