Detecting Cyberbullying using Latent Semantic Indexing

Jacob L. Bigelow
Ursinus College
Collegeville, PA USA
jabigelow@ursinus.edu

April Edwards (Kontostathis)
Ursinus College
Collegeville, PA USA
april.edwards@ursinus.edu

Lynne Edwards
Ursinus College
Collegeville, PA USA
ledwards@ursinus.edu

ABSTRACT

Cyberbullying has proven consequential to youth Internet users and previous methods relied heavily on the use of manually developed dictionaries. This project describes preliminary results for a system that uses Latent Semantic Indexing (LSI) for the detection of cyberbullying in a labeled collection of posts from Formspring.me. After preprocessing to account for variations in spelling and use of emoticons, a search system was developed. Our system significantly outperforms the baseline with a very simple query and is not dependent on a dictionary of bullying terms.

CCS Concepts

•Information systems → Web and social media search; *Data stream mining;* **•Security and privacy → Usability in security and privacy;**

Keywords

Cybersafety; Cyberbullying Detection; Latent Semantic Indexing

1. INTRODUCTION

According to a survey done by the Cyberbullying Research Center more than 80% of teens use a cell phone, making them highly susceptible to cyberbullying [2]. Cyberbullying is harassment done through the use of technology and could be anything from nasty cellphone text (SMS) messages to posting hurtful messages on social networking sites. Another survey found that one in three young people have experienced cyber threats while using the Internet [2].

Using a collection of 13,159 posts from the online social media site Formspring.me, we have developed a cyberbullying retrieval process using Latent Semantic Indexing (LSI), which is said to bring out the 'latent semantics' in a collection of documents. As described in section 3, LSI relies on a development of a term by document matrix, which relies on a set of standard terms. Because Formspring.me is an online forum, the site is littered with spelling errors, word spelling variations (sometimes used to avoid filtering or parental control software), and emoticons (sequences of punctuation that is used to convey emotion), the original dataset contained almost 40,000 unique words. In section 5 we describe our process for cleansing the data, reducing the term count to 5,716. The Formspring.me collection, described more fully in section 4, contains 848 cyberbullying interactions, a density of 6.4%. In section 6 we show how our process is able to significantly improve over this baseline, using a simple query and without relying on a standard dictionary of bullying terms.

CyberSafety'16 October 28-28 2016, Indianapolis, IN, USA

ACM ISBN 978-1-4503-4650-4/16/10.

DOI: http://dx.doi.org/10.1145/3002137.3002144

2. RELATED WORK

Within the research community, there is a growing focus on the development of algorithms and tools to automatic detect cyberbullying on social media platforms. In [8], the authors focus on the prevalance of different features in cyberbullying content, and offer an analysis of the social relationships between users. [3] also focuses on user information as an supplement to content analysis only. In 2012, Dinakar, et al. developed an approach to combating the cyberbullying problem. The technique is dependent on a specialized knowledge base of terms [5]. Hosseinmardi et al. have done an analysis of the user from Ask.fm, another question and answer website [7].

To our knowledge, only two papers have attempted to use latent semantic analysis on this problem. Xu, et al. take a contextual approach, looking at both the content of the original post and at the response [13], and Kontostathis, et al. use LSI to identify the most prominent bullying terms in Formspring text [10].

3. BACKGROUND

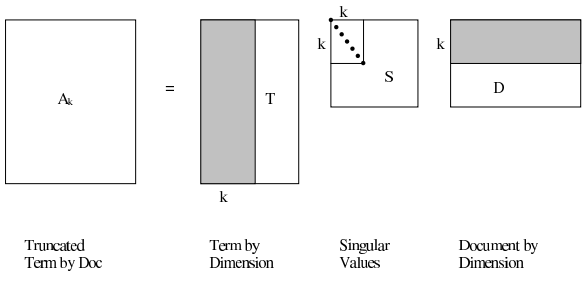
This section offers a brief tutorial on LSI, which is based on traditional vector space retrieval.

3.1 Traditional Vector Space Retrieval

In traditional vector space retrieval, documents and queries are represented as vectors in t-dimensional space, where t is the number of indexed terms in the collection. Generally the document vectors are formed when the index for the collection is generated. These vectors form a matrix that is often referred to as the term-by-document matrix, A. The query vector, q, is formed when a search of the collection is performed.

In order to determine the relevance of a document to the query, the query vector is multiplied by the term-by-

Figure 1: Truncation of SVD for LSI



document matrix and a results vector, w, is computed as shown in equation 1. The score in position j of this vector provides a measure of the similarity of document d_j to q. Generally a search and retrieval system will order the scores in descending order, and return the documents to the user in this order. This provides a *ranking* of the documents for each query. The document with the highest score is given $rank = 1$.

$$w = qA \tag{1}$$

The entry in the term by document matrix can be simple or complex. In these preliminary studies, a simple term count was used.

3.2 Latent Semantic Indexing

LSI is based on a mathematical technique called Singular Value Decomposition (SVD). The SVD process decomposes a term-by-document matrix, A, into three matrices: a term-by-dimension matrix, T, a singular-value matrix, S, and a document-by-dimension matrix, D. The number of dimensions is r, the rank of A. The original matrix can be obtained, through matrix multiplication of TSD^T. This decomposition is shown in equation 2.

$$A = TSD^T \tag{2}$$

In an LSI system, the T, S and D matrices are truncated to k dimensions. This truncation is accomplished by removing columns $k+1$ to r of T, columns and rows $k+1$ to r of S, and $k+1$ to r of D^T. This process is shown graphically in Figure 1 (taken from [1] – the shaded areas are kept). Dimensionality reduction is thought to reduce 'noise' in the term-by-document matrix, resulting in a richer word relationship structure that many researchers claim reveals latent semantics present in the collection [4, 6]. Queries are represented in the reduced space by:

$$qT_k$$

where T_k is the term-by-dimension matrix, after truncation to k dimensions. Queries are compared to the reduced document vectors, scaled by the singular-values ($S_k D_k^T$). This process provides a similarity score for each document for a given query. Thus, the truncated term-by-document matrix is shown in equation 3, and the result vector, w, is produced using equation 4. As with vector space retrieval, the scores will be sorted in descending order and the system will return documents to the user in rank order.

$$A_k = T_k S_k D_k^T \tag{3}$$

$$w = qA_k \tag{4}$$

Choosing an optimal dimensionality reduction parameter (k) for each collection remains elusive. Traditionally, the optimal k has been chosen by running a set of queries with known relevant document sets for multiple values of k. The k that results in the best retrieval performance is chosen as the optimal k for each collection. Optimal k values are typically in the range of 100-300 dimensions [6, 9]. For our experiments, we set $k = 100$, after determining that larger or smaller k values did not significantly alter the results.

4. FORMSPRING.ME DATASET

The website Formspring.me is a question-and-answer based website where users openly invite others to ask and answer questions. What makes this site especially prone to cyberbullying is the option for anonymity. Formspring.me allows users to post questions anonymously to any other user's page. Some instances of bullying found on Formspring.me include: "Q: Your face is nasty. A: your just jealous" and "Q: youre one of the ugliest bitches Ive ever fucking seen. A: have you seen your face lately because if you had you wouldn't be talkin hun (:." It is interesting to note that the reactionary tone of the answers lends weight to the labeling of these interactions as containing bullying content. As noted in [13] sometimes bullying is identified by a defensive or aggressive response.

To obtain this data, we crawled a subset of the Formspring.me site and extracted information from the pages of 18,554 users. The XML files that were created from the crawl ranged in size from 1 post to over 1000 posts. For each user we collected the following profile information: date the page was created, userID, name, link(s) to other sites, location, and biography.

The name, links and biography data were manually entered by the user who created the page (the Formspring.me account) so we cannot verify the validity of the information in those fields. In addition to the profile information, we collected the following information from each Question/Answer interaction: Asker UserID, Asker Formspring.me page, Question, and Answer.

We extracted the question text and the answer text from a randomly chosen subset of the Formspring.me data and used Amazon's Mechanical Turk service to determine the labels for our corpus. Mechanical Turk is an online marketplace that allows requestors to post tasks (called HITs) which are then completed by workers. The "turkers" are paid by the requestors per HIT completed. The process is anonymous (the requestor cannot identify the workers who answered a particular task unless the worker chooses to reveal him/herself). The amount offered per HIT is typically small. We paid three unique workers 5 cents to label each post. Each HIT we posted displayed a question and its corresponding answer from the Formspring.me crawl.

The primary advantage to using Mechanical Turk is that it is quick. Our dataset was labeled within hours. We asked three workers to label each post because the identification of cyberbullying is a subjective task. We used a voting system for determining cyberbullying content. At least two workers had to identify the post as cyberbullying for it to be

recorded as a positive instance in our system. Of the 13,159 posts that were submitted for labeling, 848 were judged to contain cyberbullying content (a hit rate of 6.4%, which is consist with prior research on Formspring data [12]). A full discussion of the development and labeling of the dataset appears in [10].

5. DEVELOPING THE TERM LIST

A large number of the terms in the raw text are either misspelled, are online abbreviations (lol, hahah), or are emoticons (sequences of characters meant to depict emotion or intent). Before forming the term by document matrix, we preprocessed the data set to eliminate these anomalies, whenever possible.

Common emoticons such as smiley faces (:), :], :D), frowny faces (:(, :[ ), faces that are winking (;)), heart symbols (<3), and faces with a tongue out(:p, :P) were replaced by text equivalents ('smileyface', etc.). All remaining punctuation was removed, and then one character words were removed.

To identify internet abbreviations, we check a list of common internet terms, such as hahah or wht, and acronyms, such as idk or lol. If the post word matches something on this list, it is either left alone (with the internet spelling) or converted to the more common abbreviation of the word.

LanguageTool, an open source proof reading program [11], was used to correct spelling errors in the remaining terms. The software comes with a standard English dictionary and also allows for the addition of words to ignore. The user can choose different rules to guide the tool on the appropriate ways to resolve misspelled words. Of course this approach causes some words to be incorrectly changed, but overall was useful for improving our term list and reducing the term list size.

Initially when creating the term list we did not reduce terms that were in all caps to lower case, under the assumption that words in all caps may have a particularly strong relationship to cyberbullying. Interestingly, this decision had a detrimental effect on the results, and we reverted to the traditional approach, converting all text to lowercase.

Finally we removed all remaining terms with a global term count greater than 1000 or less than 3. After preprocessing, we reduced the original term count from approximately 40,000 unique terms to 5716 unique terms, a reduction of 85%.

6. RESULTS

A very simple query was used initially to test the system. The query was "you dirty, ugly, piece of shit. I hope you die." This was designed to test the power of LSI - to determine if the system would be able to find the latent semantics, even with a very short post, that could not contain all of the keywords. The first match returned was "Q: hi asslee asshole :D A: Hi anonymous" which does not contain any of the query keywords.

Figure 2 shows the precision results for this query for selected ranks. Precision at rank n is calculated as the number of true positives in the top n documents (posts) returned from the query system, divided by n. Precision in this case peaks at 55%, and remains at or above 40% through the top ranked posts. This is remarkable, given the short post length, the very short query (which was chosen almost randomly), and also remembering that that baseline is 6.4%. For top ranks we are doing far better than chance – 7 to 9 times the baseline. Figure 3 show the trend when we continue precision measurements from 100 to 400 (10-100 is included for comparison). Although we see a drop off, we are still doing significantly better than the baseline.

Figure 2: Precision at top ranks with short query

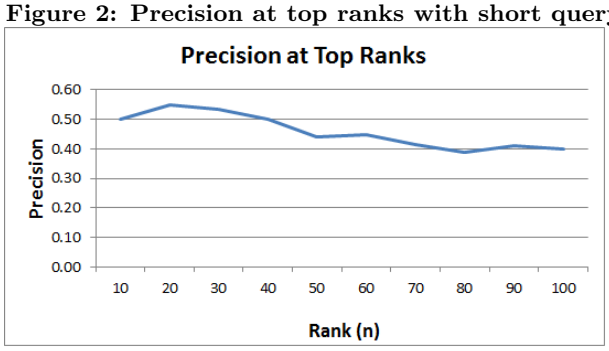
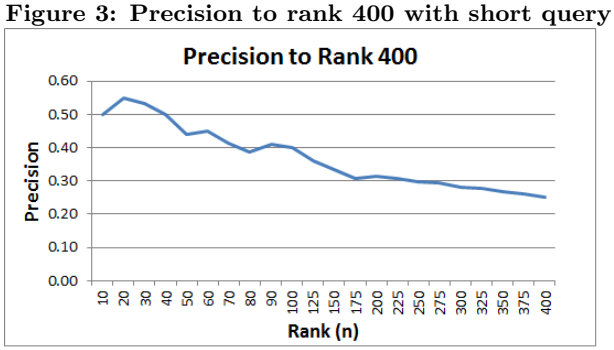


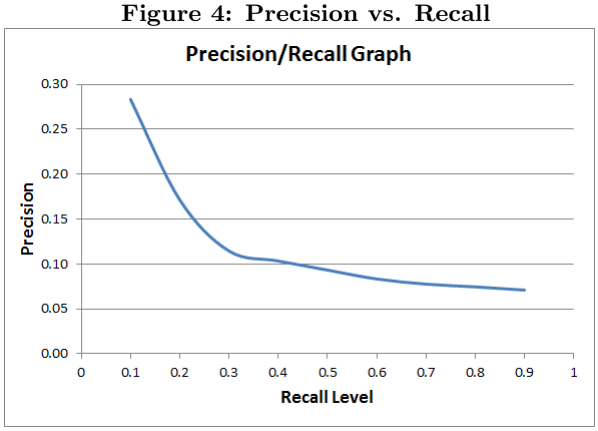
Figure 3: Precision to rank 400 with short query



To get a better sense of the performance, as rank increases, we measured precision at various recall levels (shown in Figure 4). The graph shows a significant decline after recall .1, indicating that we are pushing some of the cyberbullying posts into the top ranks, but are missing many of them.

We expected the results to improve when we ran the ex-

Figure 4: Precision vs. Recall



periments with a much longer query - a query comprised of all of the terms in all of the cyberbullying posts. Instead the performance degraded to below the baseline. We found only one cyberbullying post in the top 40 returned documents (posts), a precision of .025. Considering that selecting records at random would have produced precision of .064, this was surprising. Past results imply that longer queries would usually improve performance [10]. This may be due to the term weighting scheme (simple term count) that was applied to both the posts and the query, or it may be because the query length is, in generally, overwhelming the latent semantic system. Although more research is needed to determine the exact cause, it is clear that a short, imperfect query is vastly superior to a query that, at first glance, would appear to be overfitted to the corpus.

7. CONCLUSIONS

We have described a cyberbullying detection system that is based on LSI, a well-known information retrieval technique that is said to bring out the 'latent semantics' in a collection of documents. LSI has typically been applied to longer documents containing well-formed text (for example, abstracts of scientific articles). The system we have developed detect cyberbullying in short posts that are littered with mispellings, appreviations and unusual punctuation. The system is not dependent on a database of bullying term, and, therefore, relies on LSI for semantic analysis. Preliminary results are very promising - the LSI system detected cyberbullying posts at seven to nine times the baseline.

In future work, the system will be optimized. For example, it would be interesting to see how different term weighting schemes might affect the retrieval system. It is also important to test our system on other domains, such as a collection of SMS messages or Twitter posts.

8. ACKNOWLEDGMENTS

This material is based upon work supported in part by the National Science Foundation under Grant Nos. 0916152 and 1421896. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

9. REFERENCES

[1] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):575–595, 1995.

[2] bullyingstatistics.com. Bullying statistics. http://www.bullyingstatistics.org/content/cyber-bullying-statistics.html. Accessed: 2016-07-24.

[3] M. Dadvar and F. de Jong. Cyberbullying detection: A step toward a safer internet yard. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 121–126, New York, NY, USA, 2012. ACM.

[4] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[5] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Trans. Interact. Intell. Syst.*, 2(3):18:1–18:30, Sept. 2012.

[6] S. T. Dumais. LSI meets TREC: A status report. In D. Harman, editor, *The First Text REtrieval Conference (TREC-1), National Institute of Standards and Technology Special Publication 500-207*, pages 137–152, 1992.

[7] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra. Towards understanding cyberbullying behavior in a semi-anonymous social network. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 244–252. IEEE, 2014.

[8] Q. Huang, V. K. Singh, and P. K. Atrey. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, SAM '14, pages 3–6, New York, NY, USA, 2014. ACM.

[9] A. Kontostathis. Essential dimensions of latent semantic indexing (LSI). In *40th Hawaii International International Conference on Systems Science (HICSS-40 2007), CD-ROM / Abstracts Proceedings, 3-6 January 2007, Waikoloa, Big Island, HI, USA*, page 73, 2007.

[10] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards. Detecting cyberbullying: Query terms and techniques. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 195–204, New York, NY, USA, 2013. ACM.

[11] languagetool.org. Language tool. https://www.languagetool.org. Accessed: 2016-07-24.

[12] K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. In *Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops - Volume 02*, ICMLA '11, pages 241–244, Washington, DC, USA, 2011. IEEE Computer Society.

[13] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore. 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, 2012. Association for Computational Linguistics.