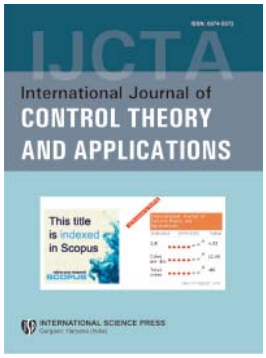

International Journal of Control Theory and Applications

ISSN : 0974–5572

Volume 9 • Number 42 • 2016

A Novel Hybrid Approach for Noise Removal from Short Text

NavneetKaur[a] and Navdeep Kumar[a]

[a]*Department of Computer Science and Engineering, Chandigarh University, India*

E-mail: navneet.cu.2016@gmail.com, navdeep.cb@gmail.com

Abstract: This paper contains the study of noise removal techniques from short text user-generated data on internet. The dataset of twitter has been considered and noisy twitter tweets are normalized as per requirement. The out of vocabulary and in vocabulary words are retrieved from the user generated text. The accuracy is measured. The proposed hybrid technique outperforms the existing techniques. Large number of experiments has been performed for application domain of text normalization.

Keywords: *Text normalization, noise removal, hashtag word segmentation.*

1. INTRODUCTION

Social media data is unstructured and ill-formed data which is available on internet.Online Social Networks (OSNs) are becoming an appealing source of information for the users as they can share their ideas,opinions and contents in lesser amount of time and cost.There are many OSNs available these days such as Twitter, Facebook, Myspace, LinkedIn etc. Among these Twitter is becoming more popular day by day.Many organizations such as news agencies share their news and messages using Twitter.With the increase in its popularity, the attacks on Twitter are also increasing because among millions of users not all the users are legitimate or genuine.

Twitter has rapidly come out as an attractive social network where millions of people share and discuss about events [14], news,opinions etc. [4][7][10][15][19] Although the structure of Twitter messages contains only 140 characters but it rapidly emerged as a system where real time information can be obtained as users can post and receive messages to and from their followers immediately and the information can be disseminated to even more users. Due to these services like spreading news, posts, events, ideas etc. Twitter is becoming an open opportunity for different types of spammers. These include:

Phishers: Phishers are those user accounts which steal user's information like password, credit card information etc. They spread wrong URLs in the tweets and the users who click on these links enters into the wrong websites through which their information is stolen.

Marketers: Marketers focusses on disseminating the advertisements for the products in order to popularize them. These may not harm the user but can mislead the legitimate users.

Adult Content Propagators: These spammers flood their tweets with adult content which redirects the user to the malicious sites.

Malware Propagators: Malware is a malicious software which is harmful for the legitimate or genuine users. Spammers add malicious links in the tweets which leads to automatically downloading of the malwares.

On Twitter, the textual messages posted by the users are known as tweets and each tweet contains only 140 characters. Therefore, most of the spammers add short URLs in their tweets to mislead the users. There are many URL shortener available such as Twitter URL shortener, Bitly, Google URL shortener etc. Also, Twitter has some features for better user interaction which are influenced by the spammers. Noise removal and text normalization [1][2][3] are similar concepts which provide useful text of the given text. This useful text can be further processed for multiple applications. The hashtag normalization and at meantion normalization are key to find named entities and topic regarding which the text is. Thereafter, no slang dictionary is used to synchronize the text for the same.

As preventive measure,rules against spam and abuse on Twitter have been released.The accounts that will violate the rules results in permanent suspension of that account.The rules set divides the spam Twitter on the basis of behavior,content and social relationship.But spammers coordinate multiple accounts so that their aim to manipulate the users is achieved and they are not detected by Twitter rules.

2. LITERATURE SURVEY

Bo Han et al. [9] have worked on lexical normalization of short text messages to detect the ill-formed words(misspellings, abbreviations etc.) and normalize them in a standard form. The dataset used is SMS corpus and a novel Twitter dataset of 549 English tweets. The proposed system consists of lexical normalization and candidate selection process [15-19]. The Stanford Parser for the extraction of dependency based features from NYT (New York Times) corpus is used. On the basis of these features a linear kernel SVM classifier on Twitter clean data is trained. Then ill-formed words are extracted from OOV words for normalization and the most likely candidate from the confusion set is selected.The system is evaluated with Precision,Recall and F-score metrics with 61.1%,85.3% and 71.2% respectively and results that the proposed approach performs better for Twitter messages than SMS.Infuture,improvement of ill-formed word detection method should be done as some of the ill-formed words are meaningless or irrelevant.

Max Kaufmann[11] has worked on syntactic normalization of Twitter tweets to normalize the tweetsand to convert them into standard form of English.The dataset used is Edinburg Twitter Corpus of 97 million tweets out of which 1 million tweets are extracted.This approach follows two steps namely Pre-processing and Machine Translation. In pre-processing. Orthographic Normalization and Syntactic Normalization is done to remove the orthographic errors and elements like @username and #(topic name).In Machine Translation,tweets are trained using SMT tool Moses and then translation is done to get the normalized tweets. BLEU and NIST scores gives the results on system performance before and after the normalization of tweets and concludes that BLEU scores are significantly affected with 18% increase in them. In future, summarization of Twitter messages should be done to achieve the more summarized and useful messages with better accuracy.

In previous research, hybrid technique for all noise removal techniques is not implemented [5][6][9][12] [15][16]. In this research, we plan to implement the hybrid technique to remove noise from text.

3. PROPOSED METHODOLOGY

The proposed methodology as mentioned in Fig 1, shows that different attributes are extracted from each tweet.

The proposed technique undergoes two phases namely calculation of in-volcabulary words (IV words) and out of vocabulary words (OOV words) before and after noise removal. The user generated short text contains lot of un-related information. The information which is required is hidden or is written in ill-formed language. This language is human readable but in-correct. This language contains slangs, abbreviations etc. At initial stage the IV and OOV words are calculated by checking every word using dictionary and Wikipedia. If it is available in English dictionary or at Wikipedia, then the text is considered to be valid and is thus called in-vocabulary text

of IV word. The words which are out of vocabulary may or may not be useful text. This may the case where the text may be either abstract or new named entities might have been introduced. The idea in this research is to normalize the text in such a way that the user generated text may be in readable form.

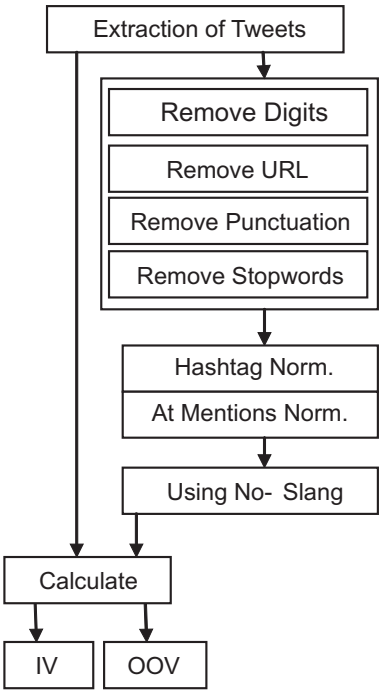

Figure 1: Architecture of proposed technique

In the beginning, the pre-processing is done using removal techniques. Punctuation marks are removed because any grammatical meaning using punctuation or any emoticon is not required in the text during normalization. The text is normalized by considering important terminology which is required to make text topic specific. URLs and digits are also of no use as important information is not communicated via URLs or digits. Finally, it has been observed that stop words are of no use in indicating important information and hence, removed.

Next, it has been observed that hashtag normalization is required. The hashtag normalization is done by separating the words as per their meanings if written in well-formed format. For instance #HashTag is separated as 'Hash Tag'. Similarly '#Hotrod_coupe' is separated as 'Hotrod coupe'. In the same way, at mentions are normalized. Finally, the text is converted into lowercase to remove the ambiguity between words like 'Hotrod' and 'hotrod'. Thus, same text is considered as similar words.

Finally, www.noslang.com is used to remove slangs from the text. The slangs in the text are replaced by full forms as mentioned at website. There are thousands of slangs which are available in user-generated text as mention at website. The request is send to the website and full form for slang abbreviation is obtained. Normalized Hashtags, at-mentions and full forms are replaced in existing text. Finally, IV and OOV are obtained.

4. RESULTS AND DISCUSSION

The results thus obtained are mentioned in table 1. This table shows that out of vocabulary words and in-vocabulary words are thus obtained before and after text normalization. In table 2, it has been observed that as the size of dataset is increased, the accuracy obtained for the data is improved. This is probably due to the fact that ill-formed words which are left are very few in large datasets as compared to smaller ones because as the size gets increased, the words get repeated and slangs or Hashtags thus normalized are more appropriate. Finally, fig 2., shows that the proposed hybrid technique performs the best results.

Table 4 shows the comparison of results as obtained using different techniques. It has been observed that the proposed hybrid approach outperforms the existing techniques. Total number of words are increased using hybrid technique that other two techniques because of the fact that slangs like lol gives three words instead of one 'laughing out loud'. Similarly, #Hashtag word normalization gives multiple words instead of one. Thus, hybrid technique gives better performance with 91.53% than existing techniques.

Table 1
Results of proposed methodology for noise removal

Number of Tweets	*Dataset*	*OOV – B*	*IV – B*	*OOV – A*	*IV – A*
10	Dataset 1	37	110	15	135
37	Dataset 2	108	467	48	519
50	Dataset 3	45	431	37	459
1443	Dataset 4	1198	14768	492	15988

Table 2
Results of proposed methodology for noise removal

Dataset	*Accuracy Before*	*Accuracy After*
Dataset 1	74.82%	90%
Dataset 2	81.12%	91.53%
Dataset 3	90.54%	92.54%
Dataset 4	92.49%	97.01%

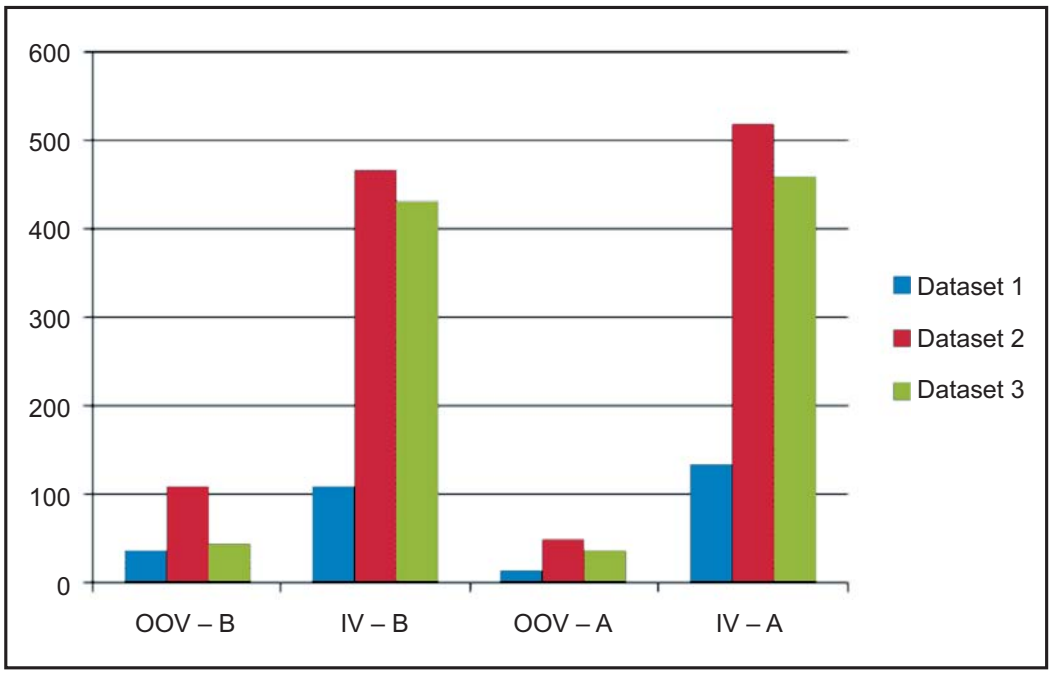

Figure 2: Improved In-vocabulary words obtained

Table 2
Comparison of results of proposed methodology for noise removal

	M1: Noslang	*M2: Hashtag*	*M3: Hybrid*
OOV – B	108	108	108
IV – B	467	467	467
OOV – A	50	50	48
IV – A	459	435	519
Accuracy	90.17%	89.69%	91.53%

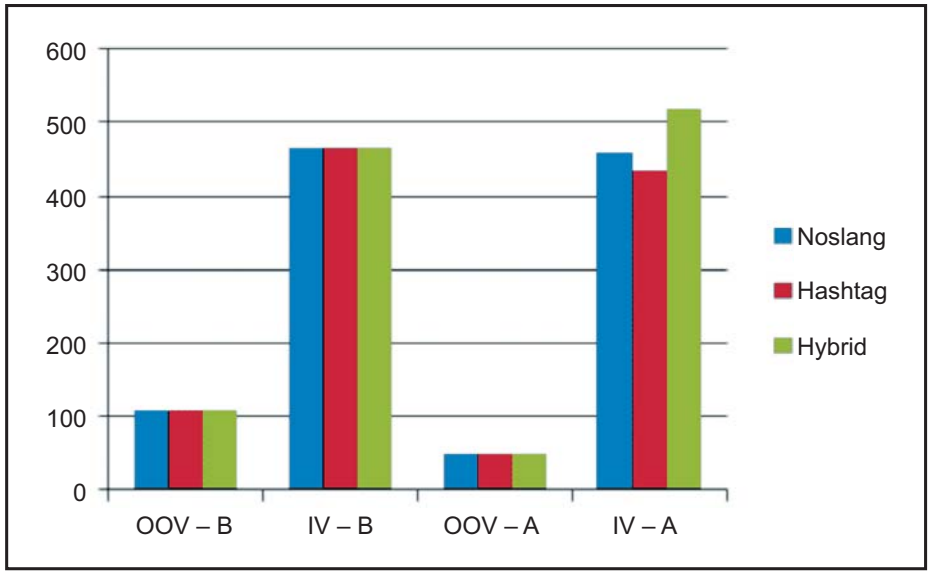

Figure 3: Comparison of results as obtained for different techniques

5. CONCLUSION

This research paper discusses the study of different noise removal techniques required for spam detection application domain. The accuracy obtained with proposed technique is improved as size of dataset is increased.

REFERENCES

[1] Aw, A., Zhang, M., Xiao, J., & Su, J. (2006, July). A phrase-based statistical model for SMS text normalization. In Proceedings of the COLING/ACL on Main conference poster sessions (pp. 33-40). Association for Computational Linguistics.

[2] Baldwin, T., & Li, Y. An In-depth Analysis of the Effect of Text Normalization in Social Media. 2015

[3] Baldwin, T., Kim, Y. B., de Marneffe, M. C., Ritter, A., Han, B., &Xu, W. (2015). Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *ACL-IJCNLP 2015*, 126.

[4] Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., &Basu, A. (2007). Investigation and modeling of the structure of texting language.International Journal of Document Analysis and Recognition (IJDAR), 10(3-4), 157-174.

[5] Chrupała, G. (2014). Normalizing tweets with edit scripts and recurrent neural embeddings. In Proceedings of ACL.

[6] Cook, P., & Stevenson, S. (2009, June). An unsupervised model for text message normalization. In Proceedings of the workshop on computational approaches to linguistic creativity (pp. 71-78). Association for Computational Linguistics.

[7] Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., ...& Smith, N. A. (2011, June). Part-of-speech tagging for twitter: Annotation, features, and experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 (pp. 42-47). Association for Computational Linguistics.

[8] Han, B., & Baldwin, T. (2011, June). Lexical normalisation of short text messages: Maknsens a# twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 368-378). Association for Computational Linguistics.

[9] Han, B., Cook, P., & Baldwin, T. (2012, July). Automatically constructing a normalisation dictionary for microblogs. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (pp. 421-432). Association for Computational Linguistics.

[10] Hassan, H., &Menezes, A. (2013, August). Social Text Normalization using Contextual Graph Random Walks. In ACL (1) (pp. 1577-1586).

[11] Kaufmann, M., &Kalita, J. (2010, July). Syntactic normalization of twitter messages. In International conference on natural language processing, Kharagpur, India.

[12] Kobus, C., Yvon, F., &Damnati, G. (2008, August). Normalizing SMS: are two metaphors better than one?. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 (pp. 441-448). Association for Computational Linguistics.

[13] Liu, F., Weng, F., & Jiang, X. (2012, July). A broad-coverage normalization system for social media language. In Proceedings of the 50th Annual Meeting of the Association for uComputational Linguistics: Long Papers-Volume 1 (pp. 1035-1044). Association for Computational Linguistics.

[14] Muskan& Kumar M. Review on Event Detection Techniques in social multimedia. *Online Information Review.* Vol. 40, issue 3. Emerald Insight. 2016

[15] Porta, J., & Sancho, J. L. (2013). Word Normalization in Twitter Using Finite-state Transducers. Tweet-Norm@ SEPLN, 1086, 49-53.

[16] Saloot, M. A., Idris, N., Shuib, L., Raj, R. G., & Aw, A. (2015). Toward Tweets Normalization Using Maximum Entropy. ACL-IJCNLP 2015, 19.

[17] Sonmez, C., &Ozgur, A. (2014). A Graph-based Approach for Contextual Text Normalization. In EMNLP (pp. 313-324).

[18] Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., & Richards, C. (2001). Normalization of non-standard words. Computer Speech & Language, 15(3), 287-333.

[19] Yang, Y., & Eisenstein, J. (2013, October). A Log-Linear Model for Unsupervised Text Normalization. In EMNLP (pp. 61-72).