Improving Web 2.0 Opinion Mining Systems Using Text Normalisation Techniques

Alejandro Mosquera
University of Alicante
amosquera@dlsi.ua.es

Paloma Moreda
University of Alicante
moreda@dlsi.ua.es

Abstract

A basic task in opinion mining deals with determining the overall polarity orientation of a document about some topic. This has several applications such as detecting consumer opinions in on-line product reviews or increasing the effectiveness of social media marketing campaigns. However, the informal features of Web 2.0 texts can affect the performance of automated opinion mining tools. These are usually short and noisy texts with presence of slang, emoticons and lexical variants which make more difficult to extract contextual and semantic information. In this paper we demonstrate that the use of lexical normalisation techniques can be used to enhance polarity detection results by replacing informal lexical variants with their canonical version. We have carried out several polarity classification experiments using English texts from different Web 2.0 genres and we have obtained the best result with microblogs where normalisation contribution to the classification model can be up to 6.4%.

1 Introduction

Nowadays, Web 2.0 applications provide some of the most popular forms of communication between users on the Internet such as blogging, social networks or short text messaging platforms. This large daily amount of generated information contains valuable insights about user opinions and sentiments regarding almost any topic.

A basic task in opinion mining deals with determining the overall polarity orientation of a document about some topic. The polarity information extracted from user comments and consumer feedback from on-line product reviews can be used to increase the effectiveness of social media marketing campaigns, discover new market threats and opportunities or react faster to customer issues. Also, microblogging platforms such as Twitter include rich metadata about interactions which provides a way to measure the reputation of their users based on the number of followers or the publication popularity by counting the number of times one message has been shared.

However, the language used in social media is very informal, containing elements such as misspellings, slang, lexical variants, inconsistent punctuations, URLs or emoticons (Thurlow, 2003). Also, the presence of genre-specific terminology such as, *RT* for *re-tweet* and *#hashtags* can make any Natural Language Processing (NLP) task challenging. For this reason, a way to handle such challenges is needed in order to automatically understand the opinions and sentiments that people are communicating on the Internet. The use of lexical normalisation techniques has recently been the subject of research applied to short and noisy texts such as tweets or SMS, improving the performance of NLP tools that need to extract contextual and semantic information from this type of informal texts.

Moreover, not all Web 2.0 genres have the same level of informality, microblog posts have to be short so they tend to contain SMS-style contractions while blog entries are usually larger and more elaborated (Santini, 2006). For this reason, in this study we evaluate the contribution of text normalisation techniques to an opinion mining application using corpora from three different Web 2.0 genres, demonstrating that it can enhance the polarity classification of microblogs by a 6.4%.

This article is organised as follows: In Section 2 we review the state of the art. Section 3 describes the normalisation process. The polarity classification is explained in Section 4. In Section 5, the obtained results are analysed. Finally, our main

Proceedings of Recent Advances in Natural Language Processing, pages 491–495,
Hissar, Bulgaria, 7-13 September 2013.

conclusions and future work are drawn in Section 6.

2 Related Work

Both academic researches and commercial companies have increased their interests recently in mining user opinions on the Internet. After the initial works of (Pang et al., 2002) several applications of opinion mining have been developed in order to measure the word of mouth (Jansen et al., 2009), correlate polls with user opinion (Balasubramanyan et al., 2010) or predicting elections results (Tumasjan et al., 2010). Most of these studies have been focused on Twitter (Barbosa and Feng, 2010), (Bifet and Frank, 2010) using both machine learning (Turney, 2002) and lexicon-based approaches (Taboada et al., 2011). The real-time nature of tweets provides a large amount of metadata and content information such as hashtags and smileys (Davidov et al., 2010) that can be used as a training corpus for opinion mining systems (Pak and Paroubek, 2010) without requiring annotated corpora (Wiebe et al., 2005).

Text normalisation techniques (Liu et al., 2011), (Han et al., 2013) based on the substitution of out of vocabulary (OOV) words have been used in opinion mining systems before (Mukherjee et al., 2012), (Gutiérrez et al., 2013), (Sidorov et al., 2013) but this process is usually presented as an intermediate filtering step without explicitly detailing the contribution of normalisation to the classification results. On the other hand, there are different genres within the Web 2.0 and they do not have the same level of informality (Mosquera and Moreda, 2012), so the contribution of text normalisation techniques to polarity classification can be more or less relevant depending on that level. For this reason, in this study we evaluate the performance of an automated opinion classification system before and after using lexical normalisation techniques using annotated corpora from three different Web 2.0 genres.

3 Lexical Normalisation

We have used TENOR (Mosquera et al., 2012), a multilingual lexical normalisation tool for English and Spanish texts in order to transform noisy and informal words into their canonical form (see Table 1). After this step they can be easily processed by NLP tools and applications.

In order to do this, OOV words are detected with a dictionary lookup. TENOR uses a custom-made lexicon built over the expanded Aspell dictionary and then augmented with domain-specific knowledge from the Spell Checking Oriented Word Lists (SCOWL)[1] package.

The OOV words are matched against a phone lattice using the double metaphone algorithm (Philips, 2000) to obtain a list of substitution candidates. With the Gestalt pattern matching algorithm (Ratcliff and Metzener, 1988) a string similarity score is calculated between the OOV word and its candidate list.

Nevertheless, there are acronyms and abbreviated forms that can not be detected properly with phonetic indexing techniques *(lol - laugh out loud)*. For this reason, TENOR uses an exception dictionary with common Internet abbreviations and slang collected from online sources[2].

Moreover, a number transliteration lookup table and several heuristics such as word-lengthening compression, emoticon translation and simple case restoration are applied to improve the normalisation results. Finally, TENOR uses a trigram language model in order to enhance the clean candidate selection.

4 Polarity Classification

The methodology explained in (Boldrini et al., 2009) has been used in order to create a two-class polarity classifier (positive/negative) based on the bag of words model. We have tested two different machine learning algorithms: j48 (Quinlan, 1993) and Support Vector Machines (SVM) (Vapnik, 1997) with word unigrams and lemmas. Stopwords and URLs were replaced by static labels with aim to simplify the model and avoid extra noise.

4.1 Datasets

We have trained our polarity classification system using annotated English texts from three different Web 2.0 genres:

Microblog publications: Extracted from 5513 Twitter messages [3].

Blog posts: The Kyoto sub-set of the EmotiBlog[4]

[1] http://wordlist.sourceforge.net/

[2] http://en.wiktionary.org/wiki/Appendix:English_internet_slang

[3] http://www.sananalytics.com/lab/twitter-sentiment/sanders-twitter-0.2.zip

[4] The EmotiBlog corpus is composed by blog posts on the Kyoto Protocol and election processes in Zimbabwe and USA

	Informal English text	Normalised English text
a)	Gotta buy this asap	I am going to buy this as soon as possible.
b)	I will nevooor buy tis again :(	I will never buy this again I'm sad.
c)	Greeeeeeat product!!!	Great product!
d)	I dnt wnt to cmplain but reply me plz	I do not want to complain but reply me please.

Table 1: Example of raw and normalised pairs of English Web 2.0 texts.

corpus.

Product reviews: The phones sub-set of the EmotiBlog[5] corpus

5 Results

The polarity classification system has been evaluated using a ten-fold cross validation, see Table 2, before and after applying normalisation techniques. Analysing the results we can observe that the overall best classification is achieved using the SVM algorithm. For the microblog genre there are improvements when using the dataset normalised with TENOR and these are higher when using lemmas instead of unigrams with a 6.4% improvement over the original dataset. However, there is almost no improvement or even the performance is decreased in some cases when applying TENOR to both blogs and reviews genres. These contain a very low amount of lexical variants and misspellings, and because of that using lexical normalisation techniques can lead to false positives in the substitution process, thus decreasing the performance of the polarity classifier. On the other hand, the microblog dataset is substantially more informal, so the application of normalisation techniques has a positive impact in the classification results.

Corpus	Method	Precision	Recall	F1
micro	j48-unigram	0.705	0.706	0.705
micro-norm	j48-unigram	0.729	0.716	0.71
blog	j48-unigram	0.836	0.88	0.828
blog-norm	j48-unigram	0.8	0.866	0.812
review	j48-unigram	0.522	0.524	0.52
review-norm	j48-unigram	0.56	0.558	0.556
micro	j48-lemma	0.637	0.626	0.626
micro-norm	j48-lemma	0.656	0.655	0.654
blog	j48-lemma	0.836	0.88	0.828
blog-norm	j48-lemma	0.803	0.861	0.817
review	j48-lemma	0.554	0.555	0.547
review-norm	j48-lemma	0.551	0.55	0.549
micro	svm-unigram	0.804	0.795	0.795
micro-norm	svm-unigram	0.83	0.83	0.83
blog	svm-unigram	**0.849**	**0.88**	**0.854**
blog-norm	svm-unigram	0.803	0.848	0.819
review	svm-unigram	**0.679**	**0.679**	**0.679**
review-norm	svm-unigram	0.662	0.662	0.662
micro	svm-lemma	0.812	0.806	0.806
micro-norm	svm-lemma	**0.858**	**0.858**	**0.858**
blog	svm-lemma	0.847	0.877	0.854
blog-norm	svm-lemma	0.79	0.837	0.809
review	svm-lemma	0.667	0.667	0.667
review-norm	svm-lemma	0.671	0.671	0.671

Table 2: Polarity classification results before and after normalisation using corpora from three different Web 2.0 genres.

6 Conclusions and Future Work

In this paper we have presented the evaluation of the contribution of a text normalisation tool to an opinion mining system using corpora from three different Web 2.0 genres. The application of lexical normalisation techniques to short and very noisy texts such as tweets obtained a relatively 6.4% better F1 results than the classification baseline. On the other hand, it has been shown the need to determine the level of informality before applying normalisation techniques in order to avoid the loss of information when dealing with less informal genres. The use of informality analysis and exploring different polarity classification systems are left to a future work.

Acknowledgments

This paper has been partially supported by Ministerio de Ciencia e Innovacion - Spanish Government (grant no. TIN2009-13391-C04-01), and Conselleria d'Educacion - Generalitat Valenciana (grant no. PROMETEO/2009/119, ACOMP/2010/286 and ACOMP/2011/001)

References

Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls : Linking text sentiment to public opinion time series.

[5] It is an EmotiBlog extension with reviews of mobiles phones

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Albert Bifet and Eibe Frank. 2010. Sentiment knowledge discovery in twitter streaming data. In *Proceedings of the 13th international conference on Discovery science*, DS'10, pages 1–15, Berlin, Heidelberg. Springer-Verlag.

Ester Boldrini, Javi Fernández, José M Gómez, and Patricio Martínez-Barco. 2009. Machine learning techniques for automatic opinion detection in non-traditional textual genres. *Proceedings of WOMSA*.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yoan Gutiérrez, Andy González, Roger Pérez, José I. Abreu, Antonio Fernández Orquín, Alejandro Mosquera, Andrés Montoyo, Rafael Muñoz, and Franc Cámara. 2013. Umcc_dlsi-(sa): Using a ranking algorithm and informal features to solve sentiment analysis in twitter. *Semeval 2013, Proceedings of the 7th International Workshop on Semantic Evaluations.*

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Trans. Intell. Syst. Technol.*, 4(1):5:1–5:27, February.

Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188, November.

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. 2011. Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 71–76, Portland, Oregon, USA, June. Association for Computational Linguistics.

Alejandro Mosquera and Paloma Moreda. 2012. The study of informality as a framework for evaluating the normalisation of web 2.0 texts. In *Proceedings of 17th International conference on Applications of Natural Language Processing to Information Systems (NLDB 2012)*. Springer.

Alejandro Mosquera, Elena Lloret, and Paloma Moreda. 2012. Towards facilitating the accessibility of web 2.0 texts through text normalisation. In *Proceedings of the LREC workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA) ; Istanbul, Turkey.*, pages 9–14.

Subhabrata Mukherjee, Akshat Malu, Balamurali A.R., and Pushpak Bhattacharyya. 2012. Twisent: a multistage system for analyzing sentiment in twitter. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 2531–2534, New York, NY, USA. ACM.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lawrence Philips. 2000. The double metaphone search algorithm. *C/C++ Users Journal*, 18:38–43, June.

Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

John W. Ratcliff and David E. Metzener. 1988. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46–72, July.

Marina Santini. 2006. Web pages, text types, and linguistic features: Some issues. *ICAME Journal*, 30.

Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Treviño, and Juan Gordon. 2013. Empirical study of machine learning based approach for opinion mining in tweets. In *Proceedings of the 11th Mexican international conference on Advances in Artificial Intelligence - Volume Part I*, MICAI'12, pages 1–14, Berlin, Heidelberg. Springer-Verlag.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, June.

Crispin Thurlow. 2003. Generation txt? the sociolinguistics of young people's text-messaging. *Discourse Analysis Online*, 1(1).

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM'10*, pages –1–1.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vladimir Vapnik. 1997. The support vector method. In Wulfram Gerstner, Alain Germond, Martin Hasler, and Jean-Daniel Nicoud, editors, *ICANN*, volume 1327 of *Lecture Notes in Computer Science*, pages 263–271. Springer.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210.