Using sentence-level classifiers for cross-domain sentiment analysis

Peter Kwantes
DRDC – Toronto Research Centre

Jihn Hamm
Ohio State University

Simon Dennis
University of Newcastle

Defence Research and Development Canada

Scientific Report
DRDC-RDDC-2014-R104
September 2014

Abstract

Defence Research and Development Canada has been developing a suite of capabilities built around models of semantics and visual analytic tools for Applied Research Project (ARP) 15ah. Recently, we implemented a sentiment analyser in a document visualization tool called Handles to allow users to examine the positive and negative opinions associated with concepts. The results were unimpressive. Specifically, the system does poorly classifying document from domains that are different from the training domain. In the work reported here, we consider and explore the two solutions. First we explore whether a more fine-grained analysis of sentiment where the sentences of a document are used as the functional unit of analysis rather than the whole document improves performance. Second, we increased the granularity of the classification during training from binary (positive or negative) to trinary (positive, negative, or neutral) to see if performance improved. Neither solution worked well. However, when we mixed documents from different domains together during training, we did find that the performance improved. We take the results to suggest that the best way to build a sentiment classifier that is agnostic with respect to domain is to train the classifier on examples from as many domains as possible.

Significance to defence and security

The principle purpose of an Influence campaign is to shape the attitudes and behaviour of a target audience. Over the past twenty years, increasingly sophisticated computerised methods for automatically monitoring target audience opinion and sentiment have been developed. When applied to the never-ending stream of text data generated in social media, for example, we have a potentially rich information environment in which to provide near real-time monitoring of local sentiment before, during and after an influence campaign.

Generally speaking, sentiment classifiers are trained on documents that have been pre-labelled as carrying positive or negative sentiment. After training, the algorithm applies what it has learned about sentiment to classify new documents as being positive or negative. The problem is that when the algorithm is trained on a different kind of document than those to which it is applied, it does not do a good job classifying new documents. So, for example, if we train the system to recognize sentiments by processing movie reviews, it will then do a poor job classifying product reviews, because the two classes of document use different kinds of language to express sentiment. In this paper, we explored a set of potential solutions to the domain problem that could be used in future sentiment analysis tools to improve the Canadian Armed Forces' capability to classify documents from various domains. In the end, it appears that the best way to build a classifier that is agnostic with respect to domain is to train the classifier with documents with as many diverse domains as possible.

Résumé

Recherche et développement pour la défense Canada met au point, dans le cadre du projet de recherche appliquée (PRA) 15ah, un ensemble de capacités qui reposent sur des modèles sémantiques et des outils analytiques visuels. Nous avons récemment mis en œuvre un analyseur de sentiment dans un outil de visualisation de document appelé Handles, afin de permettre aux utilisateurs d'examiner les opinions positives et négatives associées aux concepts. Les résultats n'ont pas été marquants. Plus précisément, le système ne classe pas convenablement les documents provenant des domaines qui diffèrent du domaine de formation. Nous explorons les deux solutions dans les travaux présentés ici. Tout d'abord, nous tentons de déterminer si une analyse plus poussée des sentiments, lorsqu'on emploie des phrases d'un document en tant qu'unité fonctionnelle d'une analyse plutôt qu'un document en entier, améliorerait le rendement. Ensuite, nous raffinons davantage la classification de la formation en passant d'une classification binaire (positif ou négatif) à une classification trinaire (positif, négatif ou neutre) afin de voir s'il y a eu amélioration. Aucune de ces solutions ne s'est révélée fructueuse. Par contre, lorsque nous avons regroupé, pendant la formation, différents documents provenant de divers domaines, nous avons constaté une amélioration du rendement. Les résultats de ce dernier essai nous poussent à croire qu'afin de fabriquer un analyseur de sentiment indépendant en ce qui concerne le domaine, nous devons former l'analyseur à partir d'exemples provenant d'autant de domaines possibles.

Importance pour la défense et la sécurité

Une campagne d'influence a pour objectif principal de former les attitudes et le comportement d'un groupe cible. Au cours des vingt dernières années, des méthodes informatiques de plus en plus sophistiquées ont été élaborées en vue de surveiller automatiquement les opinions et les sentiments de ces groupes cibles. Lorsque ces méthodes sont appliquées à un flot continu de données textuelles générées par les médias sociaux, par exemple, cela engendre un environnement potentiellement riche en renseignements qui nous permet d'assurer un suivi en temps quasi réel des sentiments locaux, et ce, avant, pendant et après la campagne d'influence.

En général, les classificateurs de sentiment sont formés à partir de documents qui ont été pré-étiquetés comme étant chargés de sentiments négatifs ou positifs. Or, lorsqu'un algorithme est formé à partir de types de documents différents de ceux auxquels il s'applique, ce dernier ne classifie pas les nouveaux documents adéquatement. Par exemple, si nous formons le système à reconnaître les sentiments en traitant des critiques de films, le système classifiera mal les critiques de produits, car les deux catégories de documents emploient un langage différent pour exprimer les sentiments. Nous examinons, dans le présent document, un ensemble de solutions possibles au problème des domaines, qui pourraient être employées dans le cadre de futurs outils d'analyse des sentiments afin d'améliorer la capacité des Forces armées canadiennes à classer les documents de divers domaines. En terminant, afin de fabriquer un classificateur indépendant en ce qui concerne le domaine, il semble qu'il faille former le classificateur à partir de documents provenant d'autant de domaines diversifiés que possible.

Table of contents

List of tables

1 Introduction

DRDC has been developing a suite of capabilities built around models of semantics and visual analytic tools for Applied Research Project (ARP) 15ah. Recently, we implemented a sentiment analyser in a document visualization tool called, Handles to allow users to examine the positive and negative opinions associated with concepts. Our initial findings on the classifier's performance were mixed. On the one hand, we demonstrated that we could get the system to classify documents expressing positive and negative sentiment with over 90% accuracy. On the other hand, to get that level of accuracy, the system needed to be tested on documents from the same domain used to train it. In the example we reported, the system was trained on movie reviews from the Internet Movie Database (IMDB). To get accuracy that high out of the system, it also had to be tested on movie reviews. That the training and application domains need to match is problematic because, in practice, the sentiment classifier would be intended for multiple domains.

In the work reported here, we consider and explore the solutions for two potential problems with our sentiment classifier (and classifiers more generally). First, document classifiers are trained to recognize documents that are either positive or negative in their expression of sentiment. We postulate that at least some of the inaccuracies committed by classifiers arise from the presence of sentences in the document that contain neither positive nor negative sentiment—that is, neutral sentences. Several sentences of a document carry little information about the sentiment being expressed by the content. For example, in a movie review, much of the content is designed to give the reader background and context about the film before the author starts discussing what he or she did or did not like. Sentences that contain such information will generally be neutral in tone, and as such do little to contribute to the sentiment being expressed by the document as a whole. We suspect that the presence of neutral sentiment in a document misleads the classification process, and that devising a classification algorithm that accounts for neutrality would improve performance. The second problem we addressed was the extent to which the algorithms' accuracy depends on the match between the domains from which the training and test documents were selected. Put simply, we found that our sentiment classifier did a good job correctly identifying the sentiment of movie reviews when it was trained to recognise sentiment from a collection of movie reviews. When tested on documents from a different domain, like consumer electronics, however, performance was poor.

We tested the hypothesis that examining documents down to the level of the sentence might improve the system's ability to provide accurate sentiment classifications across domains, making the classification algorithm more or less agnostic with respect to topic. The notion here was that many of the sentences expressing neutral sentiment would also often be ones that discuss content particular to the domain. For example, in a movie review, the neutral sentences will likely be those that discuss story, plot, acting and other movie-related topics. On the other hand, the sentences carrying sentiment however are perhaps more likely to use language that is not strongly tied to a particular topic.

1.1 The classifier

One device commonly used to classify documents into arbitrary categories is the Support Vector Machine (SVM). An SVM uses documents (potentially thousands) that have been pre-classified as, say positive or negative in sentiment as a training set. The training of the SVM occurs roughly as follows:

1. The documents of the training set are processed to identify all the unique word stems (e.g., part is the word stem for the terms parted, parting, parts). The word stems are referred to as, features.

2. Each document is transformed into a sparse vector containing the transformed frequency of the document's features across all features available in the training set.

3. A document vector can be thought of as a set of coordinates describing the document's location as a point in space. So, for example, in the same way that a 3-dimensional vector describes a point in three-dimensional, or (x, y, z) coordinates, the document vector describes the document as a point in N-dimensional space, where N is the number of unique features in the training set.

4. During training, the SVM considers the locations of all the documents/points in N-dimensional space and works out a function that differentiates one category of documents from the other. In our case, a function that best differentiates documents carrying positive sentiment from negative sentiment.

5. After training, the SVM uses the function it has derived to categorize new documents as being either positive or negative in the sentiment they express.

1.2 The experiment

In this experiment, we will consider two domains, movie reviews and electronic product reviews. We will explore the possibility that a classifier trained on sentence-level expressions of sentiment improves sentiment analysis done across domains over the more typical document-level classifier.

As mentioned above, many sentences of a document express neither positive nor negative polarities but rather are instead neutral and would not contribute to the prediction of sentiment in the corresponding documents. We postulate therefore that having a classifier ignore neutral content might improve accuracy. Second, by using features from single sentences instead of the entire document during training, we might lessen the impact that rare or domain-specific words have on predicting sentiment for the document and, as a result reduce the negative impact that mismatched domains have in cross-domain predictions.

One caveat worth mentioning is that using sentence-level classifiers reduces the amount of information being used to classify documents. There are far fewer words in a sentence than there are words in most documents and therefore fewer features upon which to distinguish positive from negative sentiment. To augment the words used in the classification, we used two kinds of feature collections to train the classifiers. In the first, more traditional fashion, we trained the

classifiers on features comprised solely of single words. In the second, we included bigrams (word pairs that occurred together more than twice in the collection) in the list of features.

Our hypothesis is that expressions of sentiment based on sentence-level information will be more agnostic with respect to domain than document-level information, and as a result, sentiment classification done across domains when sentences are aggregated will be more accurate than when document-level classification is performed across domains. Further, we hypothesize that identifying sentences in a document that carry neutral sentiment during the classifier's training might help the classifier differentiate positive and negative documents.

2 Method

2.1 Materials

2.1.1 Subjectivity lexicon

In [2], an external database – the Multi-Perspective Question Answering (MPQA) subjectivity lexicon [5][a] was used to incorporate prior information on sentiment polarity of words. The lexicon contains words and phrases that express "subjective states" ([5], p. 3) like good and evil which are clearly positive and negative, respectively as well as words that are neutral in polarity but provide clues that a sentiment's expression is forthcoming. For example, verbs like feel and think will often be followed by an expression of sentiment, and words like totally and completely are often used to express the intensity of a sentiment. The lexicon consists of 2718 positive and 4911 negative words, and was used as the basis on which sentences from the Movie Review and the Product Review were selected. We noted that the lexicon contains more negative than positive words—a bias that is not explained by its creators. Details on how it was used in the experiment reported here are provided in the following sections.

2.1.2 Movie review data

The Movie Review data [4][b] consists of 1,000 positive and 1,000 negative movie reviews crawled from the IMDB movie archive, with an average of 30 sentences in each document. We preprocessed the data. First, punctuation, numbers and other non-alphabet characters were removed. Second, for the purpose of reducing the vocabulary size and addressing the issue of data sparseness, stemming was performed using the Porter stemmer algorithm [6]. The Porter stemmer was chosen because it is the most common one used in information retrieval experiments. Stopwords were also removed based on a list [2][c]. After preprocessing, we had 633,648 words with 25,248 distinct terms. The total number of bigrams was 280,234; however, after excluding those that appeared less than three times in the corpus, 33,868 remained.

2.1.3 Product review data

The Product Review data [1][d] is a collection of product reviews from Amazon.com. It contains four types of review domains: books, DVDs, electronics and kitchen appliances. Each review is assigned a positive or negative label based on the rating score given by each reviewer. In each domain, there are 1,000 positive and 1,000 negative reviews. In this report, we only used the electronics category. We preprocessed the corpus in the same way as the Movie Review data: removing punctuation, numerals and any other non-alphabet characters, stemming words using the Porter stemmer, and removing stopwords from the same list. After preprocessing, there were 7,125 distinct words and 47,171 bigrams in the corpus. After reducing the bigrams to those which appeared at least three times in the corpus, there were 5,285.

2.2 Procedure

2.2.1 Training the classifier for document-level sentiment classification

Features for documents were treated as a vector of unigram counts or unigram and bigram counts together. Then, as described above, the vector of features for each document was used to derive the function that differentiates documents expressing positive and negative sentiment.

2.2.2 Training the classifier for sentence-level sentiment identification and applying it to document classification

For the purpose of training sentence-level classifiers, we selected one sentence from each document from the Movie Review and the Product Review datasets. For the Movie Review, we observed that the majority of the sentences had neutral sentiments, such as descriptions of plots or particular scenes from the movie. Hence, randomly selecting a sentence from any review would have resulted in most labels being neutral and would not have been useful in determining the positive and negative polarity of sentences. Instead, we used a different criterion for selecting sentences. Using the subjective lexicon, we identified every sentence in the document that contained more than four subjective words and randomly choose one of the sentences. In so doing, we increased the chance of positive or negative sentences being selected.

For the Product Reviews neutral sentences were not as frequent as they were in the Movie Reviews, but the length of sentences was generally much shorter. For this reason, the sentence we chose from the document was the one that had the largest number of words from the subjective lexicon.

For the 2,000 sentences from the Movie Reviews[e] the number of unigrams and bigrams (as subsets of those for the whole documents) were 5,785 and 6,787, respectively. For the 2,000 sentences taken from the Product Reviews[f], the number of unigrams and bigrams were 3,025 and 3,008, respectively.

To collect ground-truth sentiment labels for the 4,000 sentences, we used the consensus from workers providing judgments using Amazon's Mechanical Turk (MT). Each unprocessed sentence was read by five (sometimes ten) MT workers who were asked to judge the sentiment of the sentence as being positive, neutral, or negative in polarity. The workers were rewarded by US$0.05 per judgment, and were allowed to rate as many sentences as they wanted. Ground truth labels were determined by the raters' consensus: if the responses from the workers for a sentence were either unanimous or if the modal response contained 80% of the responses, we considered it the "true" label.

About 67% of the sentences (1322 for the Movie Review and 1344 for the Product Review) had such consensus, and we discarded the rest of the sentences in training the classifiers. We used the unigrams and bigrams from the corpus of each domain separately as features to train the classifier.

We used the linear SVMs in two ways: to make binary classifications where the sentence labels positive or negative, and to make trinary (positive vs. neutral vs. negative) classifications.

Once the sentence-level classifiers were trained, the algorithm was applied to all sentences from the domain's collection so that every sentence of every document was given a binary or trinary label. The predictions from the multiple sentences in a document were accumulated as binary or trinary votes, and used as intermediate input to a second-stage linear SVM, which produces the final positive or negative sentiment prediction of the document.

3 Results

3.1 Within-domain classification performance

In this sub-section we describe classifier performance in the case where the training domain was matched to that of the testing domain. For testing the classifier's accuracy within a domain, we used a technique called 10-fold cross-validation. Essentially, the technique involves dividing N documents of a training set into ten groups of N/10 documents. Then over ten iterations, one of the ten groups is used to test the classifier after being trained on a group of documents comprised of those in the other nine groups. The classifier's performance is taken as the average accuracy across the 10 iterations, and presented in Table 1.

Classification using document-level classifiers. Using the features from the whole text, the 10-fold cross-validation accuracy for the Movie Review documents was 0.836 when unigrams were used as features and 0.841 when they were augmented with bigrams. For the Product Reviews, accuracy was 0.8060 (unigram) and 0.8070 (unigram+bigram).

Classification using sentence-level classifiers. Recall that when the document classification is done on the basis of sentences, it occurs in two stages. In the first, the sentences for which we have ground-truth sentiment ratings are used to tag the remaining sentences in the collection with a sentiment label. How well does it work? For the experiment conducted here, when the features from sentences were used, the 10-fold cross-validation accuracy for the training sentences from the Movie Review was 0.7807 (unigram) and 0.7932 (uni+bigram) for binary classification, and 0.6213 (unigram) and 0.6138 (unigram+bigram) for the trinary classification. For the Product Review, the 10-fold cross-validation accuracy for the training sentences was 0.7614 (unigram) and 0.7183 (unigram+bigram) for binary classification, and 0.6573 (unigram) and 0.6362 (unigram+bigram) for trinary classification.

In the next stage, the sentence-level classifiers were applied to the documents and the counts of positive, negative and, when relevant, neutral sentences were accumulated, the resultant binary or trinary features were used to train the second-level classifier of sentiments at the document level. For the Movie Review, the accuracy of 10-fold cross-validation was 0.7350 (unigram) and 0.7310 (unigram+bigram) using binary features, and 0.7450 (unigram) and 0.7365 (uni+bigram) using trinary features. For the Product Review, the accuracy of 10-fold cross-validation was 0.7630 (unigram) and 0.7615 (unigram+bigram) using binary features, and 0.7665 (unigram) and 0.7735 (unigram+bigram) using trinary features.

3.2 Cross-domain classification performance

In cross-domain sentiment analysis, the classifier is trained on documents from a domain that is different from that of the documents being used during test. One issue we were forced to consider was the extent to which any of the domain-specific terms contributed to the expression of sentiment, and that if the documents from one domain had many more unique words than the other, crossing domains during test would confound domain and the size of the feature list. We reasoned that the fairest test of cross-domain classification performance would be to use only those terms and bigrams that were shared in the documents from both domains. The Movie

Review collection and Product Review collection had 25,248 and 7,125 words, respectively. For bigrams, the Movie Review collection and Product Review collection had 33,868 and 5,286, respectively. When we created a common list containing words and bigrams shared between the two collections, the features consisted of 4,897 words and 2,466 bigrams. The unigrams and bigrams for the sentence-level classifiers were also similarly restricted to the common vocabularies. The accuracy data are shown in Table 2.

Classification using document-level classifiers. When the source domain was the Product review and the target domain was the Movie Review, the accuracy of cross-domain classification using whole documents was 0.5970 (unigram) and 0.5960 (unigram+bigram). When the domains were reversed, the accuracy of cross-domain classification was 0.6835 (unigram) and 0.6605 (unigram + bigram).

Classification using sentence-level classifiers. When the source domain was the Product Reviews and applied to classifying movie reviews, the accuracy using sentence-level classifiers was 0.5705 (unigram) and 0.5310 (unigram+bigram) using binary features, and 0.5765 (unigram) and 0.5595 (unigram+bigram) using trinary features. When the order of the domains was reversed, accuracy of classification using sentence-level classifiers was 0.6045 (unigram) and 0.5865 (unigram+bigram) using binary features, and 0.6180 (unigram) and 0.5755 (unigram + bigram) using trinary features.

Building a document classifier that is somewhat insensitive to the mismatch between training and testing domain is difficult. Our strategy to institute the insensitivity was to devise means of stripping domain-relevant information and reducing its impact on the decision. Another strategy, which may hold more promise, would be to take a completely opposite approach. Instead of removing all sources of domain information, include as much domain information as possible during training. In doing so, we might increase the likelihood that the features of the test documents match at least one of the domains represented in the training documents and boost classification performance.

To test the notion, we conducted another test using the sentence data we collected from MT. We re-ran the sentence-based classifier to classify sentences as positive and negative in expressed sentiment. This time however, sentences from both domains were mixed together. Done this way, the sentence classifier is not domain-specific. We then used the newly trained classifier to label positive and negative (and neutral) sentences in the documents from both domains. As before, the positive and negative (and neutral) sentences were tallied as votes, and the resulting binary vector was used to predict sentiment on the documents from both domains. The results are shown in Table 3 alongside the reproduced performance values for sentence-based classifier performance for the within- and cross-domain predictions from Table 1 and Table 2. The results were striking. When domains were mixed during training, classification performance from either domain was equal to that of sentence-based classification when training and testing were conducted within the same domain (see Table 1) and far superior to the accuracy we reported for crossed domains in Table 2.

Table 1: *Within-domain sentiment classification accuracy for documents in both domains.*

Domain	Features	Training Material		
		Documents	Sentences	
			Pos/Neg	Pos/Neg/Neut
Movie Reviews	Unigram	0.84	0.74	0.75
	Unigrams & Bigrams	0.84	0.73	0.74
Product Reviews	Unigram	0.81	0.76	0.76
	Unigrams & Bigrams	0.81	0.77	0.77

Pos = Positive, Neg = Negative, Neut = Neutral.

Table 2: *Cross-domain sentiment classification accuracy for documents in both domains.*

Features	Training Material		
	Documents	Sentences	
		Pos/Neg	Pos/Neg/Neut
Unigram	0.64	0.59	0.60
Unigrams & Bigrams	0.63	0.56	0.57

Table 3: *Sentence-based document classification performance for within- and cross-domain document predictions.*

Classification	Training Context		
	Within-Domain	Cross-Domains	
		Mismatched	Mixed
Pos/Neg	0.75	0.56	0.75
Pos/Neg/Neut	0.76	0.57	0.76

4 Discussion

We tested two techniques to make a document classifier more agnostic with respect to the domain from which it is trained. Table 1 summarized the result of within-domain sentiment classification using document- and sentence-level features.

Table 2 summarized the result of cross-domain sentiment classification from this section collapsed across domains. Across both within-domain and cross-domain we found that augmenting unigram features with bigrams did nothing to improve performance. We take the finding to mean that, in at least our case, the number of unigram features that we extracted from the document collection was adequate to yield optimal performance of the classifier, and so adding bigrams provided no additional value. However our finding may not necessarily generalize to other document sets.

We consider next the results for document classifications done within domains. It is not entirely surprising that document-based classification was superior to sentence-based classification. In the latter case, classification is done on the pattern of sentiment judgements done by the classifier across the sentences of a document. To the extent that the classifier commits errors classifying sentences within documents, we should expect some increased inaccuracies in the classification of the documents. What is somewhat surprising, however, is that when the sentences are further classified to include the "neutral" category, performance does not improve. In the case where the classifier makes binary classifications of a document's sentences, any neutral sentence in the document will be erroneously classified as positive or negative. Not so for sentences classified using the trinary classification. By having the classifier identify neutral sentences within a document, the system should have had a more accurate representation of the distribution of positive and negative sentiment it contained. As a result, we would have expected subsequent document classification performance in the trinary condition to be better than that in the binary condition. The finding is worth further investigation in future work.

Our results for classifications done between domains were disappointing. We hypothesized that when training and test were conducted from documents in different domains, sentence-level classifiers might perform better than document-level classifiers because we could a) exclude text that contained domain specific content and b) account for the text within a document that carried no information about sentiment (i.e., neutral content). Our expectation therefore, was that cross-domain classification performance would be best in the condition where sentence-based training using trinary features was applied to the documents. Not surprisingly, cross-domain classifiers generally are much worse at classifying text than within-domain classifiers. However, contrary to our hypothesis, document-level classifiers still performed better than sentence-level classifiers. The gap in performance between the two was reduced; however it is only because document-based classification performance suffers more greatly when domains are crossed than sentence-based classification does. Also contrary to our hypothesis, sentence-level classifiers using binary (positive vs negative) features and trinary (positive vs negative vs neutral) features performed at about the same level of accuracy.

5 Recommendations and Conclusions

In this report, we explored the possible advantages of using sentence-level sentiment classifiers for the task of classifying the sentiment of new documents in a different domain. Our aim was to use the work here to fine tune the sentiment classification algorithm currently implemented in Handles, DRDC's document visualization tool. Experimental results suggested that there is still work to be done before more traditional methods for predicting sentiment can be abandoned. We recommend further work exploring how using text from multiple domains can have the desired effect of making a document classifier relatively independent of any particular domain. The other recommendation we would suggest is the exploration of other classification mechanisms. We have not, for example, tested some of the other cross-domain methods for sentiment analysis reported in the literature (e.g., [3] or [1]) which use structural or spectral alignment of features to improve cross-domain tasks.

This page intentionally left blank.

References

[1] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 432–439, 2007.

[2] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. Proceedings of the 18th ACM conference on Information and knowledge management, CIKM '09, pages 375–384, New York, NY, USA, 2009. ACM.

[3] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In Proceedings of the 19th international conference on World Wide Web, WWW '10, pages 751–760, New York, NY, USA, 2010. ACM.

[4] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the ACL, 2004.

[5] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.

[6] Porter, Martin, F. An algorithm for suffix stripping, Program, 14, pages 130-137, 1980.

[a.] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

[b.] http://www.cs.cornell.edu/people/pabo/movie-review-data/

[c.] http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words/

[d.] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

[e.] 31 sentences were discarded from analysis due to corruption.

[f.] One sentence was discarded from analysis.

List of symbols/abbreviations/acronyms/initialisms

ARP	Applied Research Program
DND	Department of National Defence
DRDC	Defence Research and Development Canada
DSTKIM	Director Science and Technology Knowledge and Information Management
IMDB	Internet Movie Database
MPQA	Multi-perspective Question Answering
MT	Mechanical Turk
SVM	Support Vector Machine

DOCUMENT CONTROL DATA

(Security markings for the title, abstract and indexing annotation must be entered when the document is Classified or Designated)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g., Centre sponsoring a contractor's report, or tasking agency, are entered in Section 8.)

DRDC – Toronto Research Centre
1133 Sheppard Avenue West
PO Box 2000
Toronto, ON M3M 3B9

2a. SECURITY MARKING (Overall security marking of the document including special supplemental markings if applicable.)

UNCLASSIFIED

2b. CONTROLLED GOODS

(NON-CONTROLLED GOODS)
DMC A
REVIEW: GCEC DECEMBER 2012

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)

Using sentence-level classifiers for cross-domain sentiment analysis

4. AUTHORS (last name, followed by initials – ranks, titles, etc., not to be used)

Kwantes, P.; Hamm, J.; Dennis, S.

5. DATE OF PUBLICATION (Month and year of publication of document.)	6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)	6b. NO. OF REFS (Total cited in document.)
September 2014	22	13

7. DESCRIPTIVE NOTES (The category of the document, e.g., technical report, technical note or memorandum. If appropriate, enter the type of report, e.g., interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Scientific Report

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

DRDC – Toronto Research Centre
1133 Sheppard Avenue West
PO Box 2000
Toronto, ON M3M 3B9

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)
15ah	W7711-088147/001/TOR

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC-RDDC-2014-R104	

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

Unlimited

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

Unlimited

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

DRDC has been developing a suite of capabilities built around models of semantics and visual analytic tools for Applied Research Project (ARP) 15ah. Recently, we implemented a sentiment analyser in a document visualization tool called Handles to allow users to examine the positive and negative opinions associated with concepts. The results were unimpressive. Specifically, the system does poorly classifying document from domains that are different from the training domain. In the work reported here, we consider and explore the two solutions. First we explore whether a more fine-grained analysis of sentiment where the sentences of a document are used as the functional unit of analysis rather than the whole document improves performance. Second, we increased the granularity of the classification during training from binary (positive or negative) to trinary (positive, negative, or neutral) to see if performance improved. Neither solution worked well. However, when we mixed documents from different domains together during training, we did find that the performance improved. We take the results to suggest that the best way to build a sentiment classifier that is agnostic with respect to domain is to train the classifier on examples from as many domains as possible.

Recherche et développement pour la défense Canada met au point, dans le cadre du projet de recherche appliquée (PRA) 15ah, un ensemble de capacités qui reposent sur des modèles sémantiques et des outils analytiques visuels. Nous avons récemment mis en œuvre un analyseur de sentiment dans un outil de visualisation de document appelé Handles, afin de permettre aux utilisateurs d'examiner les opinions positives et négatives associées aux concepts. Les résultats n'ont pas été marquants. Plus précisément, le système ne classe pas convenablement les documents provenant des domaines qui diffèrent du domaine de formation. Nous explorons les deux solutions dans les travaux présentés ici. Tout d'abord, nous tentons de déterminer si une analyse plus poussée des sentiments, lorsqu'on emploie des phrases d'un document en tant qu'unité fonctionnelle d'une analyse plutôt qu'un document en entier, améliorerait le rendement. Ensuite, nous raffinons davantage la classification de la formation en passant d'une classification binaire (positif ou négatif) à une classification trinaire (positif, négatif ou neutre) afin de voir s'il y a eu amélioration. Aucune de ces solutions ne s'est révélée fructueuse. Par contre, lorsque nous avons regroupé, pendant la formation, différents documents provenant de divers domaines, nous avons constaté une amélioration du rendement. Les résultats de ce dernier essai nous poussent à croire qu'afin de fabriquer un analyseur de sentiment indépendant en ce qui concerne le domaine, nous devons former l'analyseur à partir d'exemples provenant d'autant de domaines possibles.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g., Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

sentiment analysis