Automatic Prediction of Evidence-based Recommendations via Sentence-level Polarity Classification

Abeed Sarker , Diego Mollá-Aliod
Centre for Language Technology
Macquarie University
Sydney, NSW 2109
{abeed.sarker, diego.molla-aliod}@mq.edu.au

Cécile Paris
CSIRO – ICT Centre
Sydney, NSW 2122
cecile.paris@csiro.au

Abstract

We propose a supervised classification approach for automatically determining the polarities of medical sentences. Our polarity classification approach is context-sensitive, meaning that the same sentence may have differing polarities depending on the context. Using a set of carefully selected features, we achieve 84.7% accuracy, which is significantly better than current state-of-the-art for the polarity classification task. Our analyses and experiments on a specialised corpus indicate that automatic polarity classification of *key* sentences can be utilised to generate evidence-based recommendations.

1 Introduction

Evidence Based Medicine is a practice that requires practitioners to rely on the best available medical evidence when answering clinical queries. While this practice improves patient care in the long run, it poses a massive problem of information overload to practitioners because of the large volume of medical text available electronically (e.g., MEDLINE[1] indexes over 22 million articles). Research has shown that the act of searching for, appraising, and synthesising evidence from multiple documents generally requires more time than practitioners can devote (Ely et al., 1999). As a result, practitioners would benefit from automatic systems that help perform these tasks and generate *bottom-line recommendations*.

In this paper, we take the first steps towards the generation of bottom-line, evidence-based summaries. Our analyses reveal that the polarities of *key* sentences in medical documents can be utilised to determine final recommendations associated with a query. *Key* sentences refer to the most important sentences in a medical abstract that are associated with a posed query. In our work, we use the sentences extracted by a domain-specific, query-focused text summariser. Consider the following sentence for example:

> *A significant body of evidence supports the use of long-acting bronchodilators and inhaled corticosteroids in reducing exacerbations in patients with moderate to severe COPD.*

The sentence is taken from a medical abstract, and clearly recommends the use of *bronchodilators and inhaled corticosteroids*, which are the context interventions in this case. In other words, it has a positive polarity for this task. Since positively polarised *key* sentences generally represent the recommendations, we attempt to automatically identify the polarities of medical sentences as the first step towards generating bottom-line recommendations. We show that sentence-level polarity classification is a useful approach for generating evidence-based recommendations. We model the problem of sentence polarity classification as a binary classification problem, and we present a supervised machine learning approach to automatically classify the polarities of *key* sentences. Our classification approach is context dependent, i.e., the same sentence can have differing polarities depending on the context.

2 Related Work

Research work most closely related to ours is that by Niu *et al.* (2005; 2006). In their approach, the authors attempt to perform automatic polarity classification of medical sentences into four categories, and apply supervised machine learning to solve the classification problem. In contrast, our approach takes into account the possibility of the same sentence having multiple polarities. This can happen when multiple interventions are mentioned

[1] http://www.ncbi.nlm.nih.gov/pubmed

International Joint Conference on Natural Language Processing, pages 712–718,
Nagoya, Japan, 14-18 October 2013.

in the same sentence, with differing results associated with each intervention. Keeping the end-use of this task in mind, we model the problem as a binary classification problem. We use the approach proposed by Niu *et al*. (2005) as a benchmark approach for comparison, and also use some of the features proposed by them.

The majority of the work related to polarity classification has been carried out outside the medical domain, under various umbrella terms such as: sentiment analysis (Pang et al., 2002; Pang and Lee, 2004), semantic orientation (Turney, 2002), opinion mining (Pang and Lee, 2008), subjectivity (Lyons, 1981) and many more. All these terms refer to the general method of extracting polarity from text (Taboada et al., 2010). The pioneering work in sentiment analysis by Pang *et al*. (2002) utilised machine learning models to predict sentiments in text, and their approach showed that SVM classifiers (Vapnik, 1995) trained using bag-of-words features produced good accuracies. Following this work, such classification approaches have been applied to texts of various granularities: documents, sentences, and phrases. Research has also focused on classifying polarities relative to contexts (Wilson et al., 2009). However, only limited research has taken place on applying polarity classification techniques on complex domains such as the medical domain (Niu et al., 2005; Sarker et al., 2011).

Our aim is to investigate the possibility of using sentence-level polarity classification to generate bottom-line, evidence-based summaries. While there has been some research on automatic summarisation in this domain (Lin and Demner-Fushman, 2007; Niu et al., 2006; Sarker et al., 2013; Cao et al., 2011), to the best of our knowledge, there is no system that currently produces bottom-line, evidence-based summaries that practitioners can utilise at point of care.

3 Data, Annotation and Analysis

We use the corpus by Mollá and Santiago-Martinez (2011), which consists of 456 clinical questions, sourced form the Journal of Family Practice[2] (JFP). Each question is associated with one or more bottom-line answers (multi-document summaries) authored by contributors to JFP. Each bottom-line answer is in turn associated with detailed explanations provided by the JFP contributors; these detailed explanations are generally single-document summaries. The corpus also contains abstracts of source documents that provide the evidence of the detailed explanations.

The bottom-line summaries in the corpus present final recommendations in response to the queries. For example, a bottom-line summary may or may not recommend an intervention in response to a disorder. Thus, the bottom-line summaries can be considered to be polarised — when an intervention is recommended, the polarity is positive, and when it is not recommended, the polarity is non-positive. The bottom-line summaries are generated by synthesising information from individual documents. Therefore, it is likely that the polarities of the individual documents, or their summaries, agree with the polarities of the associated bottom-line summaries.

For the preliminary annotation and analysis, we used the same data as the task-oriented coverage analysis work described in (Sarker et al., 2012). The data consists of 33 manually identified questions. All these questions are treatment questions and the bottom-line summaries mention one or more interventions, some of which are recommended while the others are not. We first annotated the polarities of the bottom-line answers relative to the interventions mentioned. We used two categories for the annotation — recommended/not recommended (positive/non-positive). Figure 1 presents a question, the associated bottom-line summary, and our contextual polarity annotation. All the answers to the 33 questions were annotated by the first two authors of this paper. In almost all the cases, there was no disagreement between the annotators; the few disagreements were resolved via discussion.

Next, we collected the *key* (summary) sentences from the abstracts associated with the bottom-line summaries. To collect the *key* sentences from the documents, we used the QSpec summariser (Sarker et al., 2013), which has been shown to generate content-rich, extractive, three-sentence summaries. We performed polarity annotation of these summary sentences. Similar to our bottom-line summary annotation process, for a sentence, we first identified the intervention(s) mentioned, and then categorised their polarities. We came across sentences where two different interventions were mentioned and the polarities associated with them were opposite. Consider the following sentence

[2] http://www.jfponline.com

Question: *What is the most effective beta-blocker for heart failure?*
Bottom-line answer: *Three beta-blockers-carvedilol, metoprolol, and bisoprolol-reduce mortality in chronic heart failure caused by left ventricular systolic dysfunction, when used in addition to diuretics and angiotensin converting enzyme (ACE) inhibitors.*
Contextual Polarities: *carvedilol – recommended; metoprolol – recommended; bisoprolol – recommended.*

Figure 1: Sample bottom-line summary and an example of polarity annotation.

fragment, for example:

> *The present study demonstrated that the combination of cimetidine with levamisole is more effective than cimetidine alone and is a highly effective therapy ...*

For this sentence, the combination therapy is recommended over monotherapy with *cimetidine*. Therefore, the polarities are: *cimetidine with levamisole* – recommended; *cimetidine alone* – not recommended. At the same time, in a number of cases, although a sentence is polarised, it does not mention an intervention. Such sentences were annotated of this paper without adding any intervention to the context. In this manner, we annotated a total of 589 sentences from the QSpec summaries associated with the 33 questions. If a sentence contained more than one intervention, we added an annotated instance for each intervention.

A subset of the QSpec sentences, 124 in total, were annotated by the second author of this paper and these annotations were used to measure agreement among the annotators. We used the Cohen's Kappa (Carletta, 1996) measure to compute inter-annotator agreement. We obtained an agreement of $\kappa = 0.85$, which can be regarded as almost perfect agreement (Landis and Koch, 1977).

Following the annotation process, we compared the annotations of the single document summary sentences with the bottom-line summary annotations. Given that a summary sentence has been annotated to be of positive polarity with an intervention in context, we first checked if the drug name (or a generalisation of it) is also mentioned in the bottom-line summary. If yes, we checked the polarity of the bottom-line summary. In this manner, we collected a total of 177 summary sentence – bottom-line summary pairs. Among these, in 169 (95.5%) cases, the annotations were of the same polarity. In the rest of the 8 cases, the QSpec summary sentence recommended a drug, but the bottom-line summary did not.

We also manually examined the 8 cases where there were disagreements. In all the cases, this was either because individual documents presented contrasting results, i.e., the positive findings of one study were negated by evidence from other studies; or because a summary sentence presented some positive outcomes, but side effects and other issues were mentioned by other summary sentences, leading to an overall negative polarity.

If automatic sentence-level polarity classification techniques are to be used for generating bottom-line summaries in a two-step summarisation process, the first step (QSpec summaries) also needs to have very good recall. The QSpec summary sentences contained 99 out of the 109 unique interventions, giving a recall of 90.8%. We examined the causes for unrecalled interventions and found that of the 10 not recalled, 4 were due to missing abstracts from the corpus, and 2 drug names were not mentioned in any of the referenced abstracts. Thus, the actual recall is 96.1%. Considering the high recall of interventions in the summary sentences, and the high agreement among the summary sentences and bottom-line summary sentences, it appears that automatic polarity classification techniques have the potential to be applied for the task of bottom-line summary generation in a two-step summarisation process.

4 Automatic Polarity Classification

We model the problem of sentence level polarity classification as a supervised classification problem. We utilise the annotated contexts in our supervised polarity classification approach by deriving features associated with those contexts. We annotated a total of 2362 *key* sentences (QSpec summaries) from the corpus (1736 non-positive and 626 positive instances). We build on the features proposed by existing research on sentence level polarity classification and introduce some context-specific and context-independent features. The following is a description of the features.

(i) Word n-grams

Our first feature set is word n-grams (n = 1 and 2) from the sentences. Cues about the polarities

of sentences are primarily provided by the lexical information in the sentences (e.g., words and phrases). We lowercase the words, remove stopwords and stem the words using the Porter stemmer (Porter, 1980). For each sentence that has an annotated context, we replace the context word(s) using the keyword '_CONTEXT_'. Furthermore, we replace the disorder terms in the sentences using the keyword '_DISORDER_'. We used the MetaMap[3] tool (Aronson, 2001) to identify broad categories of medical concepts, known as the UMLS[4] *semantic types*, and chose terms belonging to specific categories as the disorders[5].

(ii) Change Phrases

We use the Change Phrases features proposed by Niu *et al*. (2005). The intuition behind this feature set is that the polarity of an outcome is often determined by how a change happens: if a bad thing (e.g., mortality) was $reduced$, then it is a positive outcome; if a bad thing was $increased$, then the outcome is negative. This feature set attempts to capture cases when a $good/bad$ thing is $increased/decreased$. We first collected the four groups of $good$, bad, $more$, and $less$ words used by Niu *et al.* (2005). We augmented the list by adding some extra words to the list which we expected to be useful. In total, we added 37 $good$, 17 bad, 20 $more$, and 23 $less$ words. This feature set has four features: MORE-GOOD, MORE-BAD, LESS-GOOD, and LESS-BAD. The following sentence exhibits the LESS-BAD feature, indicating a positive polarity.

> *Statistically and clinically significant improvement, including a statistically significant* ***reduction*** *in* ***mortality****, has been noted in patients receiving ...*

To extract the first feature, we applied the approach by Niu *et al*. (2005): a window of four words on each side of a MORE-word in a sentence was observed. If a GOOD-word occurs in this window, then the feature MORE-GOOD is activated. The other three features were activated in a similar way. The features are represented using a binary vector with 1 indicating the presence of a feature and 0 indicating absence.

(iii) UMLS Semantic Types

We used all the UMLS *semantic types* (identified using MetaMap) present in a sentence as features. Intuitively, the occurrences of *semantic types*, such as *disease or syndrome* and *neoplastic process*, may be different in different polarity of outcomes. Overall, the UMLS provides 133 *semantic types*, and we represent this feature set using a binary vector of size 133 – with 1 indicating the presence and 0 indicating the absence of a *semantic type*.

(iv) Negations

Negations play a vital role in determining the polarity of the outcomes presented in medical sentences. To detect negations, we apply three different techniques. In our first variant, we detect the negations using the same approach as (Niu et al., 2005). In their simplistic approach, the authors use the *no* keyword as a negation word and use that for detecting negated concepts. To extract the features, all the sentences in the data set are first parsed by the Apple Pie parser[6] to get phrase information. Then, in a sentence containing the word *no*, the noun phrase containing *no* is extracted. Every word in this noun phrase except *no* itself is attached a 'NO' tag. We use a similar approach, but instead of the Apple Pie parser, we use the GENIA Dependency Parser (GDep)[7] (Sagae and Tsujii, 2007), since it has been shown to give better performance with medical text.

For the second variant, we use the negation terms mentioned in the BioScope corpus[8] (Vincze et al., 2008), and apply the same strategy as before, using the GDep parser again. For the third variant, we use the same approach using the negation terms from NegEx (Chapman et al., 2001).

(v) PIBOSO Category of Sentences

Our analysis of the QSpec summary sentences suggested that the class of a sentence may be related to the presence of polarity in the sentence. For example, a sentence classified as *Outcome* is more likely to contain a polarised statement than a sentence classified as *Background*. Therefore, we use the PIBOSO classifications of the sentences as a feature. The sentences are classified using the system proposed by Kim *et al*. (2011) into the categories: Population, Intervention, Background, Outcome, Study Design and Other.

[3] http://metamap.nlm.nih.gov/

[4] http://www.nlm.nih.gov/research/umls/

[5] Semantic types in this category: pathological function, disease or syndrome, mental or behavioral dysfunction, cell or molecular dysfunction, virus, neoplastic process, anatomic abnormality, acquired abnormality, congenital abnormality and injury or poisoning

[6] http://nlp.cs.nyu.edu/app/

[7] http://people.ict.usc.edu/~sagae/parser/gdep/

[8] http://www.inf.u-szeged.hu/rgai/bioscope

(vi) Synset Expansion

Certain terms play an important role in determining the polarity of a sentence, irrespective of context (e.g., some of the *good* and *bad* words used in the *change phrases* feature). Certain adjectives, and sometimes nouns and verbs, or their synonyms, are almost invariably associated with positive or non-positive polarities. Thus, for each adjective, noun or verb in a sentence, we use WordNet[9] to identify the synonyms of that term and add the synonymous terms, attached with the 'SYN' tag, as features.

(vii) Context Windows

This is the first of our context sensitive features. We noticed that, in a sentence, the words in the vicinity of the context-intervention may provide useful information regarding the polarity of the sentence relative to that drug. Thus, we collect the terms lying inside 3-word boundaries before and after the context-drug term(s). This feature is useful when there are direct comparisons between two interventions. We tag the words appearing before an intervention with the 'BEFORE' tag and those appearing after with the 'AFTER' tag, and use these as features.

(viii) Dependency Chains

In some cases, the terms that influence the polarity of a sentence associated with an intervention do not lie close to the intervention itself, but is connected to it via dependency relationships, and to capture them, we use the parses produced by the GDep parser. For each intervention appearing in a sentence, we identify all the terms that are connected to it via specific dependency chains using the following rule:

1. Start from the intervention and move up the dependency tree till the first VERB item the intervention is dependent on, or the ROOT.

2. Find all items dependent on the VERB item (if present) or the ROOT element.

All the terms connected to the context term(s) via this relationship are collected, tagged using the 'DEP' keyword and used as features.

(ix) Other Features

We use a number of simple binary and numeric features, which are: context-intervention position, summary sentence position, presence of modals, comparatives, and superlatives.

[9] http://wordnet.princeton.edu/

4.1 Classification, Results and Discussion

In our experiments, we use approximately 85% of our annotated data (2008 sentences) for training and the rest (354 sentences) for evaluation. We performed preliminary 10-fold cross validation experiments on the training set using a range of classifiers and found SVMs to give the best results, in agreement with existing research in this area. We use the SVM implementation provided by the Weka machine learning tool[10].

Table 1 presents the results of our polarity classification approach. The overall accuracy obtained using various feature set combinations is shown, along with the 95% confidence intervals[11], and the f-scores for the positive and non-positive classes. The first set of features shown on the table represent the features used by Niu *et al.* (2006); we consider the scores achieved by this system as the baseline scores. The second row presents the results obtained using all context-free features. It can be seen from the table that the two context-free feature sets, expanded synsets and PIBOSO categories, improve classification accuracy from 76% to 78.5%. This shows the importance of these context-free features. All three negation detection variants give statistically significant increases in accuracy compared to the baseline.

The non-positive class f-scores are much higher than the positive class f-scores. The highest f-score obtained for the positive class is 0.74, and that for the non-positive class is 0.89. This is perhaps due to the fact that the number of training examples for the latter class is more than twice to that of the positive class. We explored the effect of the size of training data on classification accuracy by performing more classification experiments. We used different sized subsets of the training set: starting from 5% of its original size, and increasing the size by 5% each time. To choose the training data for each experiment, we performed random sampling with no replacement. Figure 2 illustrates the effect of the size of the training data on classification accuracies.

As expected, classification accuracies and f-scores increase as the number of training instances increases. The increase in the f-scores for the positive class is much higher than the increase for the non-positive class f-scores. This verifies that the

[10] http://www.cs.waikato.ac.nz/ml/weka/

[11] Computed using the binom.test function of the R statistical package (http://www.r-project.org/)

Feature sets	Accuracy (%)	95% CI	Positive f-score	Non-positive f-score
i,ii,iii, and iv (Niu et al., 2006)	76.0	71.2 – 80.4	0.58	0.83
Context-free (i-vi)	78.5	73.8 – 82.8	0.64	0.85
All (Niu)	83.9	79.7 – 87.6	0.71	0.89
All (Bioscope)	84.7	80.5 – 88.9	0.74	0.89
All (NegEx)	84.5	80.2 – 88.1	0.73	0.89

Table 1: Polarity classification accuracy scores, 95% confidence intervals, and class-specific f-scores for various combinations of feature sets.

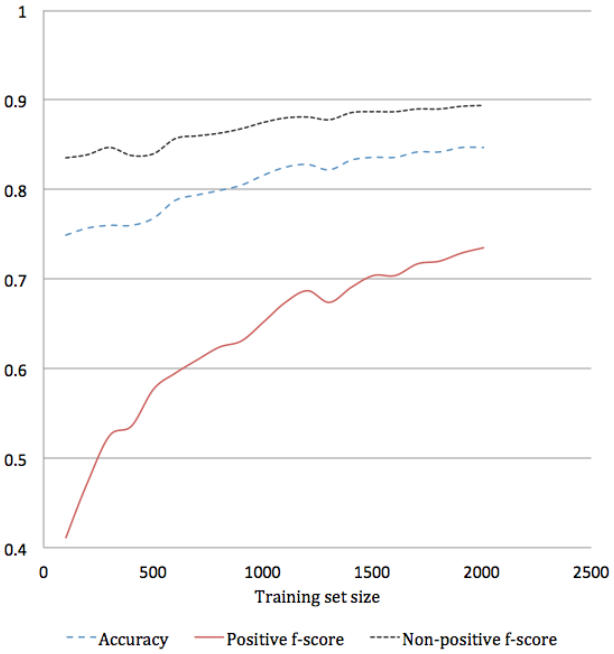

Figure 2: Classification accuracies, and positive and non-positive class f-scores for training sets of various sizes.

positive class, particularly, suffers from the lack of available training data. The increasing gradients for all three curves indicate that if more training data were available, better results could be obtained for both the classes. This is particularly true for the positive class, which is also perhaps the more important class considering our goal of generating bottom-line recommendations for clinical queries. The highest accuracy obtained by our system is 84.7%, which is significantly better than the baseline system for this domain.

To conclude this investigation, we performed manual evaluation to validate the suitability of the polarity classification approach for the generation of bottom-line recommendations. We used the 33 questions from our preliminary analysis for this. We ran 10-fold cross validation on the whole data set, collected all the sentences associated with these 33 questions, and computed the precision and recall of the automatically identified polarities of the interventions by comparing them with the annotated bottom-line recommendations. The results obtained by the automatic system were: recall - 0.62, precision - 0.82, f-score - 0.71. Understandably, the recall is low due to the small amount of training data available for the positive class, and the f-score is similar to the f-score obtained by the positive class in the polarity classification task.

5 Conclusion and Future Work

We presented an approach for automatic, context-sensitive, sentence-level polarity classification for the medical domain. Our analyses on a specialised corpus showed that individual sentence-level polarities agree strongly with the polarities of bottom-line recommendations. We showed that the same sentence can have differing polarities, depending on the context intervention. Therefore, incorporating context information in the form of features can be vital for accurate polarity classification. Our machine learning approach performs significantly better than the baseline system with an accuracy of 84.7%, and an f-score of 0.71 for the bottom-line recommendation prediction task.

Post-classification analyses showed that the most vital aspect for improving performance is the availability of training data. Research tasks specific to a specialised domain, such as the medical domain, can significantly benefit from the presence of more annotated data. Due to the promising results obtained in this paper, and the importance of this task, future research should focus on annotating more data and utilising them for improving classification accuracies. Our future research will also focus on implementing effective strategies for combining the contextual sentence-level polarities to generate bottom-line recommendations.

References

Alan R. Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. In *Proceedings of AMIA Annual Symposium*, pages 17–21.

Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont D. Antieau, Andrew Bennett, James J. Cimino, John W. Ely, and Hong Yu. 2011. AskHermes: An Online Question Answering System for Complex Clinical Querstions. *Journal of Biomedical Informatics*, 44(2):277 – 288.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. Evaluation of negation phrases in narrative clinical reports. In *Proceedings the AMIA Annual Symposium*, pages 105–109.

John W. Ely, Jerome A. Osheroff, Mark H. Ebell, George R. Bergus, Barcey T. Levy, M. Lee Chambliss, and Eric R. Evans. 1999. Analysis of questions asked by family doctors regarding patient care. *BMJ*, 319(7206):358–361, August.

Su Nam N. Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support Evidence Based Medicine. *BMC bioinformatics*, 12 Suppl 2.

J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, March.

Jimmy J. Lin and Dina Demner-Fushman. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103.

John Lyons. 1981. *Language, Meaning and Context*. Fontana, London.

Diego Mollá-Aliod and Maria Elena Santiago-Martinez. 2011. Development of a Corpus for Evidence Based Medicine Summarisation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, December.

Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. 2005. Analysis of polarity information in medical text. In *Proceedings of the AMIA Annual Symposium*, pages 570–574.

Yun Niu, Xiaodan Zhu, and Graeme Hirst. 2006. Using outcome polarity in sentence extraction for medical question-answering. In *Proceedings of the AMIA Annual Symposium*, pages 599–603.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 271–278.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, pages 1044–1050.

Abeed Sarker, Diego Mollá, and Cécile Paris. 2011. Outcome Polarity Identification of Medical Papers. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 105–114, December.

Abeed Sarker, Diego Mollá, and Cécile Paris. 2012. Towards Two-step Multi-document Summarisation for Evidence Based Medicine: A Quantitative Analysis. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 79–87, Dunedin, New Zealand, December.

Abeed Sarker, Diego Mollá, and Cécile Paris. 2013. An approach for query-focused text summarisation for evidence based medicine. In Niels Peek, Roque Marn Morales, and Mor Peleg, editors, *Artificial Intelligence in Medicine*, volume 7885 of *Lecture Notes in Computer Science*, pages 295–304. Springer Berlin Heidelberg.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2010. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307.

Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, US. Association for Computational Linguistics.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9 (Suppl 11)(S9).

Theresa Wilson, Janyce Wiebe, and Paul Hoffman. 2009. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*, 35(3):399–433.