THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Doctoral Examination Information

Report of the

New Zealand Examiner

Crn Vimiera and Pembroke Roads, Marsfield, NSW 2122
PO Box 76, Epping, NSW 1710
T (02) 9372 4704 • **ABN** 41 687 119 230

June 12th, 2017

Acquiring and Exploiting Lexical Knowledge for Twitter Sentiment Analysis
Felipe Bravo Márquez
Report by Dr Cécile Paris, FTSE. CSIRO.

The thesis addresses the issue of performing sentiment analysis on tweets in the absence of labelled data. This is an important topic: first, because the automatic classification of tweets into sentiment categories is increasingly necessary for numerous applications that use tweets as data, making this thesis very timely, and second, because manually annotating tweets to provide a training set for supervised machine learners is a very time consuming task, and, often, it is not possible to get enough labelled data. To address this issue, Mr Bravo Márquez proposes a number of methods to induce resources that can be used and evaluates them thoroughly.

The thesis is well presented, and addresses the problem under consideration in a very methodological way, with careful evaluations and analysis, and providing novel solutions. Mr Bravo Márquez explains his terminology and assumptions, introduces the various concepts, and provides ample descriptions of his new methods and their evaluations.

Mr Bravo Márquez has investigated a topic of both substance and significance, and he has done so thoroughly and critically. Throughout the manuscript, Mr Bravo Márquez demonstrates his familiarity with the relevant literature, algorithms and experimental methods and metrics. He certainly shows his knowledge of the field and of research methods. His manuscript is well written, first setting out the research aims and then methodically presenting possible solutions and their evaluations. The thesis provides indeed a comprehensive study of the solution space, with careful evaluations. The methodology employed throughout is appropriate and adequate for the topic at hand, and well applied. Mr Bravo Márquez' work constitutes a significant and substantial original contribution to the field of social media analytics, in particular sentiment classification in the Twitter environment.

Chapter 1 first introduces the thesis: the topic itself and then the methods that are being used in this work for classification: logistic regression models and Support Vector Machines (SVM). This is followed by a presentation of the research problem. In particular, Mr Bravo Márquez explains the message-level polarity problem, where a post (tweet) gets assigned a sentiment category for the post as a whole. This can be done based on the words within it, the "opinion words". Mr Bravo Márquez argues that opinion words are often domain dependent, and also may vary with time. As a result, when relying on labelled data, there is a need to obtain the appropriate labelled data sets for each domain under consideration and to constantly update this data set with new labelled examples, as time passes. This is clearly very labour-intensive and provides the motivation for Mr Bravo Márquez' work: derive polarity classifiers in label sparsity conditions. Mr Bravo Márquez proposes to solve this problem by acquiring and exploiting lexical knowledge through two specific methods: word-sentiment associations and tweet centroid models. The chapter ends with a description of the experimental framework employed in the thesis and the evaluation metrics: Precision, Recall, F1 measure, and Accuracy all fairly regular metrics, and also kappa, to rectify issues of unbalanced corpus and ROC/AUC (Receive Operating Characteristic, Area Under Curve). These metrics ensure a thorough evaluation of the work.

Chapter 2 provides an overview of prior work in sentiment classification of tweets, looking at supervised approaches (which attribute a sentiment category based on the tweet as a whole), including approaches that employ deep learning, the newest method in machine learning, and lexicon-based approaches (which looks at the specific words within the post). The chapter also looks at opinion detection (so that a sentiment label gets assigned only to tweets which express an opinion), and domain dependent and temporal issues. The chapter then turns to the acquisition of a lexicon with each word labelled as to its polarity, which can be done via a corpus-based approach or semantic networks. Mr Bravo Márquez describes a number of lexicons that have been built in prior research to support the sentiment classification task. His analysis is comprehensive, and I found section 2.4.1 very interesting and informative. The chapter continues with an overview of the applications which use tweet sentiment analysis, thereby showing the utility of the work. Finally, the chapter ends with a discussion about the interdependence of the polarity of the words in a tweet and the polarity of the tweet (which of course contains words), a relation that will be mentioned throughout the thesis, and validated later in the manuscript. In this chapter (and others), Mr Bravo Márquez shows his familiarity with the relevant literature.

Chapter 3 looks at acquiring a sentiment/opinion lexicon based on the word-sentiment association method. The method essentially combines a number of approaches: the emoticon-based approach, which annotates tweets with the sentiment expressed in the emoticon (to obtain a collection of annotated tweets); the transfer method, using tweets from another domain if the original collection does not contain many emoticons; tagging words with various features, including part-of-speech (POS) tagging; matching words with a known opinion lexicon (the “seed” lexicon) and finally training a classifier with all the available features (and labels from the seed lexicon). Through this bootstrapping method, one can obtain an expanded lexicon, thus acquiring labelled lexical knowledge. Mr Bravo Márquez suggests that this method requires a corpus in which the tweets are in chronological order to address the temporal aspect of sentiment classification. Mr Bravo Márquez tries his new method using two existing lexicons: the Stanford Sentiment Corpus (STS), an emoticon-based corpus, and the Edinburgh corpus (ED), a general corpus, which Mr Bravo Márquez then annotated using the emoticon-approach. The evaluation presented is thorough and interesting, performing both an extrinsic evaluation and an intrinsic one. We note that the task of identifying neutral words is harder than distinguishing between positive and negative. The observation that the best performance is obtained with the ensemble of expanded lexicons is noteworthy. Finally, Mr Bravo Márquez is able to identify and remove outliers caused by term ambiguity. He observes that this can cause degradations in the quality of expanded lexicons.

Chapter 4 describes the other methods Mr Bravo Márquez investigates to obtain a lexicon based on distributional models: the tweet-centroid model and word embeddings. These methods are based on the assumption that words with similar meanings (or similar sentiment polarity) occur together (or at least more often than they co-occur with words with opposing polarity). Tweet centroids are obtained by averaging all the tweets in which a word appear. Once again, a seed lexicon is used to label a subset of the words and train a classifier on the tweet centroids (which can be seen as labelled instances). The resulting classifier is then employed to classify the remaining unlabelled words. The evaluation is once again both extrinsic and intrinsic. The intrinsic evaluation is performed on the same corpora as before (STS and ED). For the positive vs negative classifier, performance varies depending on the dataset, but the use of a concatenation of vectors improves performance on both datasets. The evaluation also includes an evaluation on the task of distinguishing neutral words from sentiment-bearing words. The results suggest that world clusters are particularly useful for this scenario.

Finally, Mr Bravo Márquez compares tweet-centroids based from unigrams vs those based on a distributional representation for word semantics (PPMI). The evaluation indicates that tweet-centroids produce better vectors, and that the size of the datasets has an impact on performance. In this latter experiment, Mr Bravo Márquez also provides computational times, which is useful to understand the trade-off between performance and complexity. The extrinsic evaluation looks at the usefulness of the induced lexicon for tweet sentiment classification, and the results are quite promising, showing that the induced lexicons generally perform better than the baseline. Mr Bravo Márquez points out interesting results with respect to the size and nature of the input corpus and the various word representations. A second extrinsic evaluation looks at the lexicon themselves. There, tweet-centroids perform better than PPMI. Once again, Mr Bravo Márquez makes a few observations as to the impact of the size of the input corpus.

Next, Mr Bravo Márquez looks at expanding an emotion lexicon (so a lexicon with labels more finely defined that positive-negative), and, again, performs a thorough evaluation, both intrinsic and extrinsic. When examining the resulting lexicons, Mr Bravo Márquez notices the effect of re-tweets, which can cause an over-representation of some co-occurrences (e.g., the number 17.00 in his corpus and experiments). It would have been interesting to remove the re-tweets, re-train and re-evaluate. It is interesting that the effect of re-tweets did not come up in the earlier experiments. One might wonder, however, whether eliminating re-tweets would have produced different results. Of course, removing re-tweets from a corpus adds a step in the process, and potentially results in a much smaller corpus. It is interesting that, in the current experiment, when W2V embeddings are used, re-tweets do not seem to have much impact.

Chapter 5 looks at transferring sentiment knowledge between words and tweets, using the tweet-centroids model. Mr Bravo Márquez attempts to do this transfer in both directions: training a tweet-level polarity classifier from a polarity lexicon, and induce a polarity lexicon from a tweet polarity classifier -- thus treating words and tweets as two domains, and performing transfer learning from one domain to the other, which is useful when one domain does not have the resources required to produce a classifier. Given the "interdependence relationship" assumption between tweets and the words they contain, this transfer learning seems a plausible approach. Mr Bravo Márquez' experiments first validate this assumption, showing that sentiments of tweets are strongly related to the sentiment of the words they contain, and that the sentiment of a word is strongly related to the sentiment of the tweet in which it appears. The remainder of the chapter shows that this transfer learning is possible and yields good performances, with the quality of the induced lexicon dependent on the size and quality of the labelled Twitter data set.

Chapter 6 presents another lexicon-based distant supervision method: Annotate-Sample-Average (ASA). The goal is again to obtain a tweet sentiment classifier given an opinion lexicon. ASA generates a balanced training data set by sampling and averaging tweets containing words with the same polarity. After validating the assumption on which ASA is based (the lexical polarity hypothesis), Mr Bravo Márquez performs a number of experiments, to show that ASA can produce effective and compact training data sets for classifying tweets into positive and negative tweets. He also shows that ASA can accurately classify tweets which do not contain the words in the original source lexicon, which is interesting and useful, as it makes the method more general.

Finally, the thesis concludes with Chapter 7, with both a summary of the results and some directions for future work.

Overall, the work is novel and interesting and the thesis well written. The methods proposed by Mr Bravo Márquez are well explained, and very carefully evaluated, with extrinsic/intrinsic evaluations, qualitative and quantitative analysis, and sensitivity analysis. His induced lexicons outperform some known sentiment lexicons, and the tweet-level classifiers he obtains outperform baselines. The methods proposed by Mr Bravo Márquez are important to alleviate the problem of building sentiment classifiers in the absence of annotated data. This is frequently the case, as one studies new domains, or, as Mr Bravo Márquez points out, polarity of words can depend on time, and thus classifiers need to be retrained over time with new data. In addition, Mr Bravo Márquez is to be commanded by making all his code available to others.

I have no hesitation in allowing Mr Bravo Márquez to proceed to the oral examination.

Specific comments:

- P 18: Introduce what AUC stands for when AUC is first mentioned (at the beginning of the bullet point).
- P 50: table 2.3 It might be good to indicate in the figure which lexicon was obtained manually and which was obtained automatically (instead of having to go back to the previous page to determine which is which).
- P 73: paragraph under Table 3.6: "It is worth noting out that negative words exhibit the largest spread and that most of the boxplots show a substantial number of outliers". So? I think it would be good to add a sentence or 2 explaining the consequences of this.

- A few typos:
    - P 7, 2nd paragraph: “learning” is repeated: “supervised machine learning learning models” – I realise that this might be because the first one is for “machine learning” and the second for “learning models”, but I don’t think both are necessary.
    - P 79, to line: “the” repeated.
    - P 90: Paragraph starting with “In this study”: 3rd sentence: “The second, is a semantic…” – no comma after “second”.
    - P 130, 20th line: “the” is repeated.
    - P 143, 5th line below the figure: “improvemets” -> improvements.

Feel free to contact me if you need any further information.

Sincerely,

Dr Cécile Paris
Senior Principal Research Scientist, Research Group Leader
CSIRO
Cecile.Paris@data61.csiro.au
+61 2 9372 4704

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Doctoral Examination Information

Report of the
Overseas Examiner

Report on the PhD of Felipe Bravo Marquez entitled, "Acquiring and Exploiting Lexical Knowledge for Twitter Sentiment Analysis"

Public information to be shared with the candidate in advance of the viva

This dissertation describes a set of methods to aid the sentiment analysis of tweets, focusing on practical methods to cope with the problem of the difficulty in obtaining manually labelled tweets for machine learning purposes. Its scope includes subjectivity, polarity and emotion detection.

The goal of the thesis is appropriate and useful in the context of sentiment analysis. Although there is prior work with similar goals, this thesis contains many novel elements.

The thesis is extremely well written and presented beautifully, with substantial attention to detail. It is a pleasure to read and also very well explained at an appropriate level of depth. The explanations show a deep understanding of the relevant issues.

The literature review is appropriate in coverage and depth and with nice extra touches, such as Section 2.4.1 comparing lexicons for size and overlap.

The experiments within the thesis are carefully constructed and appropriate, giving clear and unbiased results. The use of both intrinsic and extrinsic evaluations is a particularly good idea and adds considerable weight to the findings. From the results it is possible to be confident about the value of the approaches tested.

In many places the thesis displays additional work or algorithms that show the care with which the algorithms and testing have been carried out.

The thesis includes appropriate discussions of the limitations of the approaches used, such as with emoticons not always being available or appropriate.

I only have the minor quibble that the methods chosen were not always fully justified, even though they always seemed appropriate.

The content of the thesis is clearly publishable and has been already published in good quality conferences. Conferences are more appropriate than journals in this field.

The amount and quality of the content of this thesis is above that required for a PhD and I am very happy to recommend that it proceeds to oral examination without revisions.

Questions to be shared with the candidate in advance of the viva:

Why was stochastic gradient descent used rather than any other method? (p. 67)

Did you conduct more qualitative analyses than reported in 6.4.3? This seems a bit short and could have provided more insights into the strengths and weaknesses of the method.

Following up the above question: More generally, I would have liked to see more qualitative analyses of the results in each chapter – was this something that you did but didn't report?

As a result of your work, did you get any insights into the importance of the time dimension in sentiment analysis? This was part of the problem motivation of the thesis and is included in some of the work but I did not notice any specific time-related conclusions.

Minor, optional points:

The table captions tend to be short in the early part of the thesis – more informative captions would help the reader to understand them more easily.

The title “Discussions” at the end of each chapter might scan better as just “Discussion”.

P 62, point 6: the words -> the word

P 163, second reference: capital R for O’Reilly

Prof. Mike Thelwall

University of Wolverhampton, UK

m.thelwall@wlv.ac.uk