NileULex: A Phrase and Word Level Sentiment Lexicon for Egyptian and Modern Standard Arabic

Samhaa R. El-Beltagy
Nile University
Juhayna Square, Sheikh Zayed City
Giza, Egypt
E-mail: samhaa@computer.org

Abstract

This paper presents NileULex, which is an Arabic sentiment lexicon containing close to six thousands Arabic words and compound phrases. Forty five percent of the terms and expressions in the lexicon are Egyptian or colloquial while fifty five percent are Modern Standard Arabic. The development of the presented lexicon has taken place over the past two years. While the collection of many of the terms included in the lexicon was done automatically, the actual addition of any term was done manually. One of the important criterions for adding terms to the lexicon, was that they be as unambiguous as possible. The result is a lexicon with a much higher quality than any translated variant or automatically constructed one. To demonstrate that a lexicon such as this can directly impact the task of sentiment analysis, a very basic machine learning based sentiment analyser that uses unigrams, bigrams, and lexicon based features was applied on two different Twitter datasets. The obtained results were compared to a baseline system that only uses unigrams and bigrams. The same lexicon based features were also generated using a publicly available translation of a popular sentiment lexicon. The experiments show that usage of the developed lexicon improves the results over both the baseline and the publicly available lexicon.

Keywords: Arabic sentiment analysis, Arabic sentiment lexicons, Arabic idioms, Arabic opinion mining

1. Introduction

Sentiment lexicons are an essential component of many sentiment analysis tools. Because of their importance, many sentiment lexicons for the English language, as well as for other languages, have appeared over the years. The most commonly used English lexicons include SentiWordNet (Baccianella, Esuli, and Sebastiani 2010), Ben Liu's opinion lexicon (Liu 2010), the MPQA subjectivity lexicon (Wilson, Wiebe, and Hoffmann 2005), and the NRC lexicon (Kiritchenko, Zhu, and Mohammad 2014).

Over the past five years, the topic of Arabic Sentiment analysis and opinion mining has been gaining momentum, especially since the use of the Arabic language has been increasing consistently over various social media platforms, amongst which are twitter and Facebook (Neal 2013)("Facebook Statistics by Country" 2012)(Farid 2013). However, publicly available Arabic sentiment lexicons are scarce. In fact, to the knowledge of the author the only publicly available Arabic sentiment lexicon is a translated version of the NRC word emotion association lexicon (EmoLex) (Mohammad and Turney 2013). In addition, most attempts to build Arabic lexicons have focused primarily on lexicons for Modern standard Arabic (Abdul-Mageed and Diab 2012)(Badaro et al. 2014)(Mahyouba, Siddiquia, and Dahaba 2014) (the work presented in (Al-Sabbagh and Girju 2010) is a notable exception). With the exception of EmoLex, these lexicons also do not include compound phrases and idioms. When analysing social media for sentiment, this can affect the results adversely since the language used on social media platforms, is predominantly colloquial Arabic where sentiment is sometimes expressed entirely using compound phrases or idioms (El-Beltagy and Ali 2013).

This paper presents a phrase and word level sentiment lexicon for Egyptian and modern standard Arabic, and makes it publicly available. This paper also shows that the lexicon can be useful for the task of sentiment analysis through a series of experiments.

The rest of this paper is organized as follows: section 2, provides a description of the built lexicon, section 3 presents the experiments conducted in order to evaluate the usefulness of the lexicon, and section 4 concludes this paper and presents planned future work.

2. The Lexicon

The development of the NileULex lexicon has taken place over the past two years. The first version of the lexicon, emerged from the work presented in (El-Beltagy and Ali 2013) and developed further in the work presented in (ElSahar and El-Beltagy 2014). However, manual additions and re-validations have been conducted over the past two years. The revisions were made to ensure that terms in the lexicon are of high quality and have limited or no ambiguity. For example, in the first version of the lexicon, the term "الله" was included as a positive term. The literal translation of the term "الله" is Allah or God. However, it is often used by people to express that they are in awe of something. In a way, it's the English equivalent of the word "wow". It also often precedes phrases that could be either positive or negative like for example "الله يحرقك" (may you burn in hell) or "الله يباركلك" (God bless you). To eliminate this ambiguity in the current version of the lexicon, this term has been deleted. Instead, many phrases that start with the term, were collected and added, each with its corresponding polarity. Even though, many of the terms contained in the lexicon were collected automatically, each of those entries was revised by the author. Since the author is a native speaker

of Arabic, many of the phrases contained in the lexicon were also manually collected by her from various social media postings.

The resulting lexicon contains a total of 5953 unique terms. Of those, 563 are compound negative phrases, 416 are compound positives, 3693 are single term negative words and 1281 are single term positive words. Figure 1. shows the breakdown of the lexicon. 2674 (45%) of the terms and expressions in the lexicon are Egyptian or colloquial and 3279 (55%) are Modern Standard Arabic. While most of the colloquial terms are Egyptian, a few terms from other dialects have found their way into the lexicon. Some terms that are English transliterations are also included in the lexicon, examples of which are كيوت (cute) and لايك(like). These have been included since they are commonly used in social media communication. The lexicon also includes common misspellings for a few frequently used words.

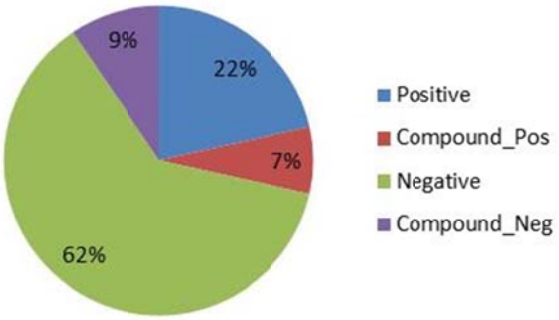

Figure 1: Class distribution within the lexicon

Term	Type	Dialect	English translation
متعدد المواهب	compound_pos	MSA	Multi-talented
ايوا بقا	compound_pos	EG	An expression denoting extreme approval
الي الجحيم	compound_neg	MSA	To Hell
مش طايقاك	compound_neg	EG	I can't stand you
جميل	positive	MSA	Beautiful
هايل	positive	EG	Excellent
اجهل	negative	MSA	The most ignorant/ I don't know
حخلع	negative	EG	I'll withdraw
ربنا يفتح عليك	compound_pos	EG	May God give you more knowledge
اتريق	negative	EG	Made fun of
بر الامان	compound_pos	MSA	Safety zone
زومبي	negative	DIA[1]	Transliterated word for 'zombie'
هاتشل	negative	EG	I am going to have a stroke
لاسع	negative	EG	Someone who has lost it
ناس بينه	compound_neg	EG	People with no class
يفوق الخيال	compound_pos	MSA	Beyond imagination

Table 1: A sample of entries in the lexicon.

Table 1 shows an example of the various entry types within the lexicon, along with their translations. Out of the seven listed compound phrases in this table, the polarity of only two entries ("بر الامان", "الي **الجحيم**") can be determined used some of their constituent words. Individually, the constituent words of the other five phrases give no indication to their polarity. This is the case with many compound phrases in the lexicon.

The common terms between the old lexicon presented in in (El-Beltagy and Ali 2013) and the presented one, represent 67.6% of NileULex.

3. Evaluation

In order to illustrate that the developed lexicon is indeed useful, it was utilized in a simple sentiment analysis task over two different twitter datasets. The datasets are described in (Khalil et al. 2015) as well as in the subsection 3.1. For carrying out sentiment analysis, a machine learning approach was followed. Minimal preprocessing was carried out on the datasets. In all experiments, input tweets are represented using the bag of words model, with uni-gram and bi-gram TF-IDF weights [16] representing the tweet vector. Lemmatization and stemming were carried out using a modified version of the stemmer presented in (El-Beltagy and Rafea 2011) which id describe in (El-Beltagy and Rafea 2016). When attempting to match input text to lexicon entries, lemmatization is preferred over stemming as the removal of a single character can change its meaning completely. If we take for example the word "روعه" which means "magnificent", and stem it using any traditional Arabic light stemmer, the result will be the term "روع", which means "terrorized". While the first term is very positive, the second is very negative. A lemmatizer on the other hand will not change the original term thus eliminating this problem.

The Weka workbench (Frank et al. 2005) was used for all described experiments. The Complement Naïve Bayes classifier (Rennie et al. 2003) with a smoothing parameter of 1, was used for classification as it demonstrated excellent performance for the task of Arabic sentiment as per the results reported in (Khalil et al. 2015). The two used datasets have three class labels (positive, negative and neutral). Two sets of experiments were carried out for each dataset. In the first, neutral instances were deleted and classification was based only on the positive and negative labels. In the second set, all instances including the neutral ones, were classified. Except for the baseline, where only uni-grams and bigrams were used as features, in each set of experiments, extra lexicon based features were introduced. Specifically, each tweet was simply annotated with the number of negative and positive lexicon entries found in the tweet (if any), and whether the last matched term was positive or negative (again, if there was a match). Compound term were given extra weight by assigning a count of 1.5 instead of 1 to each. Emoticons and other potentially useful features for the task of sentiment analysis were ignored, as the goal was to examine the impact of using a lexicon only. In addition to using the developed lexicon, Arabic entries in EmoLex (Mohammad and Turney 2013) were evaluated as well. The version of EmoLex that we have used is v0.92. This version has 5632 positive and negative entries of which 4187 are unique (1445 entries are duplicates). Arabic and other language entries in this lexicon were obtained by

[1] Dialectical word or phrase that is not specific to one Arabic speaking country or region

translating the original English version using Google translate[2].

In the final presented experiment, EmoLex and NileULex were combined to examine how an aggregate of both would affect the results. The overlap between NileULex and EmoLex is 973 terms. This represents 23.3% of the size of EmoLex and 16.5% of the size of NileULex.

3.1 The Used Datasets

The first of the two used datasets, is the Egyptian dialect dataset (NU_EG_Twitter_corpus). The tweets for this dataset were collected in December 2014 using Twitter's streaming API with Egypt's latitude and longitude as a filter. Out of the 127,000 collected tweets, 6000 tweets were randomly chosen and filtered using cosine similarity to ensure that they are all unique. The 6000 tweets were divided into six groups; 1000 tweet per group. Each group was manually annotated by three different Nile University (NU) graduate students to one of six categories namely: positive, negative, neutral, mixed, sarcastic, and ambiguous. The final annotations were selected through majority voting, with tweets exhibiting annotation conflicts removed. This dataset was then further filtered so that tweets with only positive, negative, and neutral tags, were kept. The resulting dataset contains 3436 unique tweets. To divide the dataset into training and testing sets, an eighty percent, twenty percent split was utilized. This resulted in a training set of 2746 tweets and a test set of 683 tweets, both randomly selected. The distribution of training tweets amongst polarity classes is: 1046 positive, 976 negative, and 724 neutral. The distribution of the test dataset is: 263 positive, 228 negative and 192 neutral. This dataset is available by request from the author of this paper.

The second of the used datasets is one that was collected at a research center in Saudi Arabia under the supervision of Dr. Nasser Al-Biqami who thankfully made this dataset available to Nile University for research purposes. This particular dataset, is not a public one. The tweets for this dataset were downloaded in five rounds or stages during 2014 and are mostly in Saudi dialect. The center where these tweets were collected employed a very rigorous annotation process, where multiple annotators were first trained on the annotation process and then handed subsets of the dataset to annotate. Only instances where there was an inter-annotator agreement of more than two were included in the final dataset. We have used one of the rounds for testing (1414 tweets) and the others for training (9656 tweets). The training set consists of 2686 positive, 3225 negative, and 3745 neutral tweets and the test set has 403 positive, 367 negative, and 644 neutral tweets.

3.2. Experiments and Results

Experiments over two classes

In this set of experiments, only the subset of positive and negative instances of datasets described in the previous sub-section, were used. The Egyptian dialect training subset was comprised of 2022 tweets of which 1046 were positive and 976 were negative. The test set was comprised of 495 tweets of which 254 were positive and 241 were negative. The Saudi training subset was comprised of 5893 tweets of which 2680 were positive and 3213 were negative. The test set was comprised of 768 tweets of which 402 were positive and 366 were negative. For each dataset, testing was carried out using 10 fold cross validation on the training dataset as well as using this same dataset for training and the separate test dataset for testing. The results of these experiments on each of the datasets are shown in tables 2 to 5. The best results are highlighted in the tables. The "improvement" column in the presented tables represents the percentage increase in the number of tweets correctly classified relative to the baseline.

Lexicon	10 Fold Cross Validation		
	Accuracy (%)	FScore (%)	Improvement (%)
Baseline (no lexicon)	82.59	82.6	N/A
NileULex	**85.4**	**85.4**	**3.97**
EmoLex	84.22	84.2	2.53
Combined	84.57	84.6	2.95

Table 2: 10 fold cross validation results on the 2-class Egyptian dialect dataset

Lexicon	Test Data		
	Accuracy (%)	FScore (%)	Improvement (%)
Baseline (no lexicon)	73.94	73.7	N/A
NileULex	77.58	77.6	4.92
EmoLex	76.77	76.8	3.83
Combined	**77.98**	**78**	**5.47**

Table 3: Test results on the 2-class Egyptian dialect dataset

Lexicon	10 Fold Cross Validation		
	Accuracy (%)	FScore (%)	Improvement (%)
Baseline (no lexicon)	88.7	88.7	N/A
NileULex	**89.75**	**89.7**	**1.18**
EmoLex	87.95	87.9	-0.84
Combined	89.17	89.2	0.54

Table 4: 10 fold cross validation results on the 2-class Saudi dialect dataset.

Lexicon	Test Data		
	Accuracy (%)	FScore (%)	Improvement (%)
Baseline (no lexicon)	79.17	79.1	N/A
NileULex	81.9	81.9	3.45
EmoLex	80.06	80.5	1.8
Combined	**82.3**	**82.3**	**3.95**

Table 5: Test results on the 2-class Saudi dialect dataset.

[2] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

From these results it can be seen that using both NileULex and EmoLex improve the classification accuracy in both datasets, with a notable increase when carrying out 10 fold cross validation over the Egyptian Dialect dataset. In both datasets, the combination of both lexicons seems to yield the best results over the test dataset.

Experiments over three classes

In this set of experiments, all three class labels (positive, negative, and neutral) were used. The results for each of the used datasets are presented in tables 6 through 9.

Lexicon	10 Fold Cross Validation		
	Accuracy (%)	FScore (%)	Improvement (%)
Baseline (no lexicon)	71.6	71.5	N/A
NileULex	**73.56**	**73.3**	**2.7**
EmoLex	71.52	71.3	-0.10
Combined	72.47	72.2	1.22

Table 6: 10 fold cross validation results on the 3-class Egyptian dialect dataset.

Lexicon	Test Data		
	Accuracy (%)	FScore (%)	Improvement (%)
Baseline (no lexicon)	56.22	55.7	N/A
NileULex	**59.74**	**59.4**	**6.25**
EmoLex	57.39	56.6	2.08
Combined	58.42	57.5	3.90

Table 7: Test results on the 3-class Egyptian dialect dataset.

Lexicon	10 Fold Cross Validation		
	Accuracy (%)	FScore (%)	Improvement (%)
Baseline (no lexicon)	78.88	78.9	N/A
NileULex	**79.02**	**79.0**	**0.17**
EmoLex	77.81	77.8	-1.38
Combined	78.1	78.1	-1.0

Table 8: 10 fold cross validation results on the 3-class Saudi dialect dataset.

Lexicon	Test Data		
	Accuracy (%)	FScore (%)	Improvement (%)
Baseline (no lexicon)	68.82	68.6	N/A
NileULex	**71.58**	**71.5**	**4.01**
EmoLex	70.94	70.7	3.09
Combined	71.3	71.1	3.61

Table 9: Test results on the 3-class Saudi dialect dataset.

As expected, the introduction of the neutral class causes the accuracy drops dramatically for both datasets. Unlike the experiments carried out over two classes, here the use of EmoLex actually decreases the accuracy in both datasets when carrying out 10 fold cross validation (10 FCV). The combined lexicon offers a slight improvement in the Egyptian dialect when carrying out 10 FCV, but decreases the accuracy in the Saudi dialect dataset for the same experiment. All three lexicons, enhance the results over the test dataset, with NileULex consistently providing the best enhancements.

	Training dataset (9651 tweets)		*Test dataset (1411 tweets)*	
	NileULex	EmoLex	NileULex	EmoLex
Tweets that had sentiment[3]	7312	9038	939	1311
Total matches	17023	35982	1896	5259
Unique matches	2216	2146	733	980
Total negated terms	520	797	45	101
Avg sentiWords /tweet	1.76	3.73	1.34	3.73
Tweets that had no sentiment	2339	613	472	100

Table 10: Matches between tweets and entries in the lexicon for training and testing Saudi dialect datasets.

	Training dataset (2746 tweets)		*Test dataset (683 tweets)*	
	NileULex	EmoLex	NileULex	EmoLex
Tweets that had sentiment	1879	2172	472	546
Total matches	3781	5990	903	1403
Unique matches	1324	1181	535	537
Total negated terms	148	170	39	42
Avg sentiWords /tweet	1.38	2.18	1.3	2.05
Tweets that had no sentiment	867	574	211	137

Table 11 Matches between tweets and entries in the lexicon for training and testing Egyptian dialect datasets.

Tables 10 and 11, show statistics related to the number of matches between both NileULex and EmoLex, and the used datasets. These numbers reveal that on average, more entries from EmoLex matched with tweets in the used data sets than with NileULex. Intuitively, this should have had a positive impact on the overall results of the

[3] Tweets that had at least one term matching with a lexicon entry.

experiments conducted used the EmoLex lexicon, but this was not always the case. In an attempt to understand the reason behind this, we examined a sample of words that matched with the lexicon. We discovered that some of these words were very generic and have double meanings. An example is the word: "حال", which can sometimes mean "prevented", but which more often means "situation". The fact that EmoLex was automatically translated could have easily resulted in the production of words that are not a very accurate translation of their English counter parts. In fact the lexicon comes with the following disclaimer: "some translations by Google Translate may be incorrect or they may simply be transliterations of the original English terms". We found forty eight entries in the lexicon that have both negative and positive polarity. Given the size of the lexicon, this is not a big number, but it serves to illustrate that translations are not always that accurate. Another thing we noticed, is that the volume of matches from EmoLex that were negated, was also greater than negated entries from NileULex. In our simplified sentiment analysis system, the occurrence of a negator before a sentiment term results in the reversal of its polarity. We have observed that in some cases, this is not necessarily valid. For example, the term "لا حلو", in which the negator "no" appears before the word "nice", is actually used to affirm that something is nice. We believe that proper handling of negations, elimination of noisy terms in EmoLex, and combining EmoLex with NileULex, would probably lead to better results than those presented. The goal of the experiments however, was simply to illustrate that NileULex is capable of improving sentiment analysis results even in a very simple setting.

4. Conclusion

This paper has presented NileULex, a phrase and word level sentiment lexicon for Egyptian and Modern Standard Arabic. Through a series of experiments, the presented work has shown the potential of NileULex in enhancing the results of sentiment analysis. NileULex will be available online from the LREC Resource Map and can be obtained directly from the author. We believe that it can be a valuable resource for researchers carrying out work in the area of Arabic social media sentiment analysis. Future work includes the expansion of the lexicon, and experimenting with other types of datasets. Specifically, we are currently working on assigning sentiment scores to the entries in the lexicon. Early work on this task is presented in (El-Bletagy). We are also working on a new approach for automatically learning new phrases and terms with a high degree of accuracy.

5. Bibliographical References

Abdul-Mageed, Muhammad, and Mona Diab. 2012. "Toward Building a Large-Scale Arabic Sentiment Lexicon." In *Proceedings of the 6th International Global WordNet Conference*, 18–22. Matuse, Japan.

Al-Sabbagh, R, and Roxana Girju. 2010. "Mining the Web for the Induction of a Dialectical Arabic Lexicon." In *LREC. European Language Resources …*, 288–93. http://www.lrec-conf.org/proceedings/lrec2010/pdf/344_Paper.pdf.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 0:2200–2204. doi:citeulike-article-id:9238846.

Badaro, G., R. Baly, H. Hajj, N. Habash, and W. El-Hajj. 2014. "A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining." In *Proceedings of the EMNLP Workshop on Arabic Natural Language Processing (ANLP)*, 165–73. Association for Computational Linguistics.

El-Beltagy, Samhaa R, and Ahmed Ali. 2013. "Open Issues in the Sentiment Analysis of Arabic Social Media : A Case Study." In *Proceedings of 9th the International Conference on Innovations and Information Technology (IIT2013)*. Al Ain, UAE.

El-Beltagy, Samhaa R., and Ahmed Rafea. 2011. "An Accuracy Enhanced Light Stemmer for Arabic Text." *ACM Transactions on Speech and Language Processing* 7 (2): 2–23.

El-Beltagy, Samhaa R., and Ahmed Rafea. 2016. "LemaLight: A Dictionary Based Arabic Lemmatizer and Stemmer." TR2-11-16, Technical report..

El-Bletagy, Samhaa R. "NileTMRG: Deriving Prior Polarities for Arabic Sentiment Terms." In *Proceedings of SemEval 2016 -(to Appear)*. San Diego, California.

ElSahar, Hady, and Samhaa R. El-Beltagy. 2014. "A Fully Automated Approach for Arabic Slang Lexicon Extraction from Microblogs"." *Lecture Notes in Computer Science (LNCS) - CICLing 2014, (Editor A. Glbukh)* 8403.

"Facebook Statistics by Country." 2012. http://www.socialbakers.com/facebook-statistics/.

Farid, Doaa. 2013. "Egypt Has the Largest Number of Facebook Users in the Arab World." *Daily News Egypt*, September 23. http://www.dailynewsegypt.com/2013/09/25/egypt-has-the-largest-number-of-facebook-users-in-the-arab-world-report/.

Frank, Eibe, Mark Hall, Geoffrey Holmes, Richard Kirkby, Bernhard Pfahringer, Ian H Witten, and Len Trigg. 2005. "WEKA: A Machine Learning Workbench for Data Mining." In *Data Mining and Knowledge Discovery Handbook*, edited by Oded Maimon and Lior Rokach, 1305–14. Springer. doi:10.1007/0-387-25465-X_62.

Khalil, Talaat, Amal Halaby, Muhammad H. Hammad, and Samhaa R. El-Beltagy. 2015. "Which Configuration Works Best? An Experimental Study on Supervised Arabic Twitter Sentiment Analysis." In *Proceedings of the First Conference on Arabic Computational Liguistics (ACLing 2015), Co-Located with CICLing 2015 (to Appear)*. Cairo, Egypt.

Kiritchenko, Svetlana, Xiaodan Zhu, and Saif Mohammad. 2014. "Sentiment Analysis of Short Informal Texts."

Journal of Artificial Intelligence Research 50: 723–62.

Liu, Bing. 2010. "Sentiment Analysis and Subjectivity." In *Handbook of Natural Language Processing, Second Edition*, edited by N. Indurkhya and F. J. Damerau.

Mahyouba, Fawaz H.H., Muazzam A. Siddiquia, and Mohamed Y. Dahaba. 2014. "Building an Arabic Sentiment Lexicon Using Semi-Supervised Learning." *Journal of King Saud University - Computer and Information Sciences* 26 (4): 417–24.

Mohammad, Saif, and Peter Turney. 2013. "Crowdsourcing a Word-Emotion Association Lexicon." *Computational Intelligence* 29 (3): 436–65.

Neal, Ryan W. 2013. "Twitter Usage Statistics: Which Country Has The Most Active Twitter Population?" *International Business Times*. http://www.ibtimes.com/twitter-usage-statistics-which-country-has-most-active-twitter-population-1474852.

Rennie, Jason D M, Lawrence Shih, Jaime Teevan, and David R Karger. 2003. "Tackling the Poor Assumptions of Naive Bayes Text Classifiers." *Proceedings of the Twentieth International Conference on Machine Learning (ICML)-2003)* 20 (1973): 616–23. doi:10.1186/1477-3155-8-16.

Salton, Gerard, and Chris Buckley. 1988. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing and Management* 24 (5): 513–23.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis." In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 347–54. Vancouver, Canada.