2010 IEEE International Conference on Data Mining Workshops

Mining Arabic Business Reviews

Mohamed Elhawary Mohamed Elfeky
Google Inc.
Mountain View, CA, USA
{elhawary | mgelfeky}@google.com

Abstract—**For languages with rich content over the web, business reviews are easily accessible via many known websites, e.g., Yelp.com. For languages with poor content over the web like Arabic, there are very few websites (we are actually aware of only one that is indeed unpopular) that provide business reviews. However, this does not mean that such reviews do not exist. They indeed exist unstructured in websites not originally intended for reviews, e.g., Forums and Blogs. Hence, there is a need to mine for those Arabic reviews from the web in order to provide them in the search results when a user searches for a business or a category of businesses. In this paper, we show how to extract the business reviews scattered on the web written in the Arabic language. The mined reviews are analyzed to also provide their sentiments (positive, negative or neutral). This way, we provide our users the information they need about the local businesses in the language they understand, and therefore provide a better search experience for the Middle East region, which mostly speaks Arabic.**

I. Introduction

Review websites, like Yelp.com, are popular because people like to read reviews about restaurants they consider to eat in, or shops they consider to purchase from, etc. Search engines, e.g., Google, do not provide content for the internet but their goal is to organize the content and provide the information to its users. Therefore, search engines, as well as review aggregation websites, would like to organize such business reviews and make them available when someone searches for them. Search engines do not have problems with reviews in languages with rich content over the web like English, as they can whitelist the websites that contain such reviews. The problem is in the languages with poor content like Arabic. We need to search for these reviews over the web, crawl their webpages, extract and organize them, and present them to our users when they ask for reviews. In this paper, we describe this latter system for Arabic.

The first component in this system is to determine whether an internet page contains a review or not. We have employed an in-house developed multi-label classifier that classifies any English document as a review, forum, blog, news, or shopping store. We have extended this classifier to work on Arabic documents and we used the review label to identify Arabic reviews.

Next, we need to identify the sentences in the classified-as-review page, which actually contain the review, and determine whether they contain a sentiment or not. For every sentiment, it is important to determine its polarity: positive, negative, mixed, or neutral, and its score that indicates how strong this sentiment is. This way, we can tell whether the whole review is positive or negative so that we present the information to the user sorted by the score. The sentiment analysis is also an in-house developed system [1] that classifies English sentences by polarity as positive, negative, mixed, or neutral; and by magnitude as strong or weak. We have extended this system to build an Arabic Sentiment Analyzer using MapReduce [2].

The final component of this system is how to provide the mined reviews and sentiments to the users. We annotate the documents, classified as Arabic reviews, in the search index with the extracted sentiments, and it is up to the search engine now to show those annotations as the snippet of the search result.

The remaining of the paper is organized as follows. Sections II and III describe the two above components: the reviews classifier and the sentiment analysis. In Section IV, we evaluate the performance of the system.

II. Arabic Reviews Classifier

To build the Arabic reviews classifier, we need first to collect training data from websites with Arabic language content. We collected about 2,000 URLs where about 40% of them were reviews. We found the reviews by searching the web for keywords that usually exist in user reviews like “the food was bad”, or “the bed sheets were unclean”, etc.

Next step is to determine and extract the features that can represent a review. The most important features were along the line of the number of times some keywords appear in the document. We have used the translated the features that appeared in previous classifiers for other languages and we added our own features that include Arabic keywords that usually appears in opinionated text. We have built around 1500 features that scan every piece of information we can get about a document. Fig. 1 shows a sample of the features file for Arabic. It basically describes a feature that counts the occurrences of the mentioned keywords when they have a bold typeface.

After constructing the training data, we trained an AdaBoost classifier with 200 stumps to classify whether a document is an Arabic review or not. (80% of the data were used in training and 20% in evaluation).

978-0-7695-4257-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDMW.2010.24

IEEE computer society

```
<feature>
  <type>StringTesterInSpecialText</type>
  <name>BoldReviewsOpinion</name>
  <select>bold</select>
  <CountFeatureHelper>
    <select>count</select>
  </CountFeatureHelper>
  <stringtester>
    <type>acpattern</type>
    <keyword> ••• </keyword>
    <keyword> ••••• </keyword>
    <keyword> ••••• </keyword>
    <keyword> ••• </keyword>
    <keyword> •••• ••• </keyword>
    <keyword> ••••••• ••• </keyword>
    <keyword> ••••• </keyword>
    <keyword> •••• </keyword>
    <keyword> •••• </keyword>
    <keyword> •••• ••• </keyword>
    <keyword> ••••••• •••• </keyword>
    <keyword> ••••• </keyword>
    <keyword> •••• </keyword>
    <keyword> ••••• </keyword>
    <keyword> ••••• </keyword>
    <keyword> ••••• •••• ••• </keyword>
    <keyword> ••••• </keyword>
    <keyword> •••• •• •••• </keyword>
    <keyword> ••••• </keyword>
    <keyword> •••• </keyword>
    <keyword> •••• </keyword>
    <keyword> •••• </keyword>
    <keyword> ••••• •••••••• </keyword>
  </stringtester>
</feature>
```

Figure 1. A sample features file for Arabic.

III. Arabic Sentiment Analysis

The Arabic Sentiment Analysis relies on the Arabic Lexicon (vocabulary) that contains positive and negative words / phrases ranked by their score. The standard way to generate such a Lexicon is to take a small set of manually-labeled positive and negative seed words and perform label propagation along the Arabic similarity graph.

A. Arabic Similarity Graph

The similarity graph is a graph that clusters all the words / phrases in a certain language, where two words / phrases have an edge if they are similar, i.e., they have the same polarity of sentiment, or they have the same meaning. The weight on the edge is an indication of how similar these two words / phrases are. This graph can be constructed from lexical co-occurrences in an unsupervised learning scheme from a large web corpus that can easily be trained on any language [3]. The similarity graph automatically provides extensive coverage of multi-word phrases, alternative spellings, and slang expressions.

B. Arabic Lexicon

We started by a seed list of more than 600 positive words / phrases and more than 900 negative words / phrases collected by manually looking at some of the reviews collected for the reviews classifier training; and a seed list of almost 100 neutral words / phrases, collected from the top frequent Arabic words / phrases used over the web.

We used those seed lists and the Arabic similarity graph to build the Arabic Lexicon. The Lexicon is a two column file: a phrase / word and a score. The score captures the polarity and magnitude of the phrase / word. To build the Arabic Lexicon, every word / phrase in the seed lists would get a high positive or negative score, or a neutral score in the similarity graph. Then, using label propagation, the score of each node in the similarity graph gets the sum of edges weights connecting to this node in the similarity graph, multiplied by the score of the adjacent nodes. Therefore, a node with many positive neighbors will become positive and a node with many negative neighbors will become negative. A node with a nearly balanced positive and negative neighbors will become neutral.

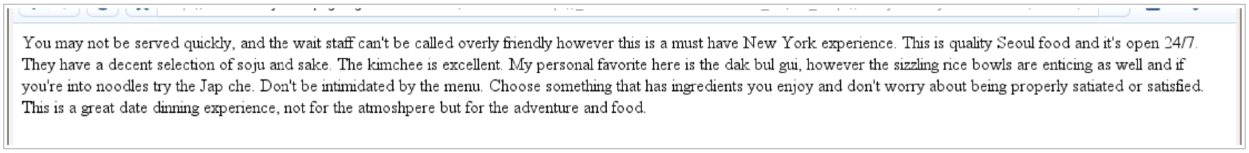
The scores of the words from the Arabic Lexicon were carefully inspected. We discovered that the top 200 words / phrases with positive polarity are not supposed to be positive at all. We traced down the positive nodes that led to such labeling through the similarity graph and we discovered that these 200 nodes shared a few garbage nodes. We discovered that, since there is sparse data coverage in Arabic, the similarity graph was built using webpages with low quality or low page rank to reach a total of 1.8 million phrases. Hence, this provides a good coverage, or so it was thought, however, we have noticed lots of garbage words in there due to the low quality of Arabic documents used. To fix this problem, we have modified the pruning scheme used to throw away nodes in the similarity graph. We came up with a new filtering rule that the words / nodes in the similarity graph with a large number of high weight edges should be thrown out from the graph. When we have applied this modification to all other languages, their corresponding Lexicon performance graphs have significantly improved.

Another important addition is that we decided to only keep the top-ranked 25 synonyms per phrase. This decision was based on the correlation between the synonym rank and accuracy used in the similarity graph: *the top 25 synonyms of a positive word tend to be >=90% positive, while all of the synonyms may only be 50-60% positive*.

The label propagation mechanism produces separate positive and negative word scores. We normalize the positive and negative scores before adding them to compensate for the negative skew (bias) in the scores. Finally, the scores are filtered by eliminating the scores below some cutoff and the log is taken. The end result of running label propagation is a Lexicon with scored words / phrases which are considered to carry positive or negative sentiments.

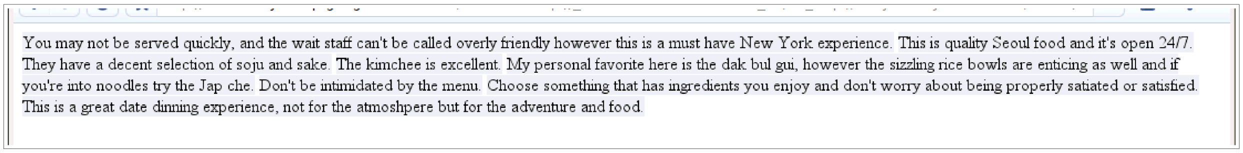
C. Negation Mechanism and Sentence Boundary Detection

A negation mechanism is needed in order to switch the polarity from negative to positive and vice versa in case there is a negation word (around 20 words in Arabic). The negation logic currently assumes that the negation term precedes the negated text, which is true for Arabic as well.



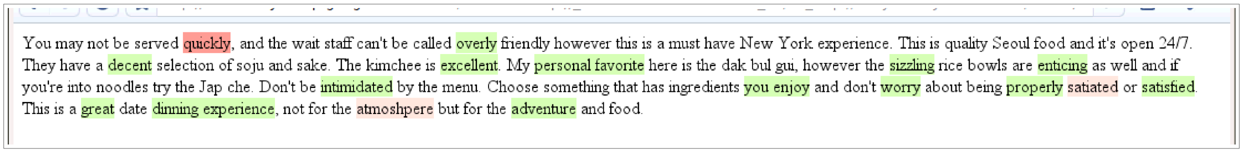
You may not be served quickly, and the wait staff can't be called overly friendly however this is a must have New York experience. This is quality Seoul food and it's open 24/7. They have a decent selection of soju and sake. The kimchee is excellent. My personal favorite here is the dak bul gui, however the sizzling rice bowls are enticing as well and if you're into noodles try the Jap che. Don't be intimidated by the menu. Choose something that has ingredients you enjoy and don't worry about being properly satiated or satisfied. This is a great date dinning experience, not for the atmoshpere but for the adventure and food.

Figure 2(a). An example English text before applying Sentiment Analysis



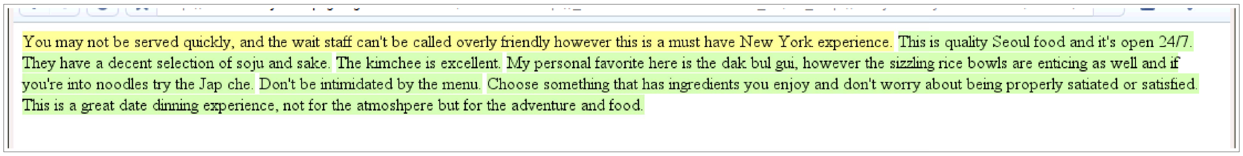
You may not be served quickly, and the wait staff can't be called overly friendly however this is a must have New York experience. This is quality Seoul food and it's open 24/7. They have a decent selection of soju and sake. The kimchee is excellent. My personal favorite here is the dak bul gui, however the sizzling rice bowls are enticing as well and if you're into noodles try the Jap che. Don't be intimidated by the menu. Choose something that has ingredients you enjoy and don't worry about being properly satiated or satisfied. This is a great date dinning experience, not for the atmoshpere but for the adventure and food.

Figure 2(b). The example above after applying sentence boundary detection; sentences are highlighted.



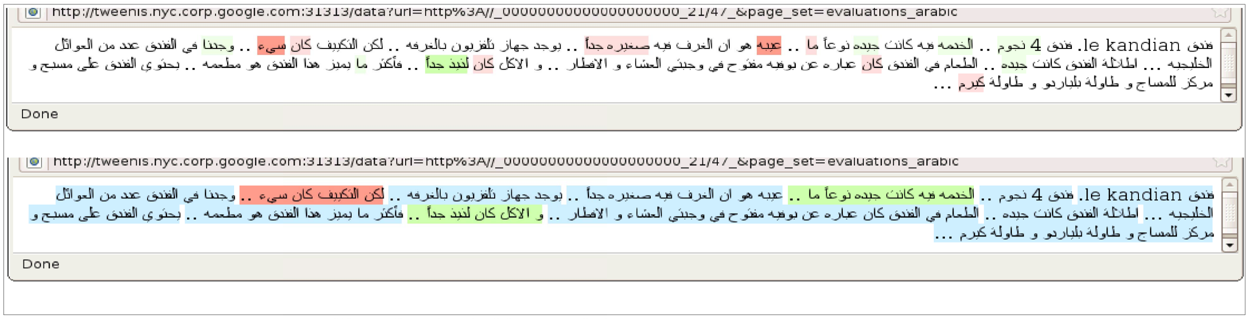
You may not be served quickly, and the wait staff can't be called overly friendly however this is a must have New York experience. This is quality Seoul food and it's open 24/7. They have a decent selection of soju and sake. The kimchee is excellent. My personal favorite here is the dak bul gui, however the sizzling rice bowls are enticing as well and if you're into noodles try the Jap che. Don't be intimidated by the menu. Choose something that has ingredients you enjoy and don't worry about being properly satiated or satisfied. This is a great date dinning experience, not for the atmoshpere but for the adventure and food.

Figure 2(c). The example above after applying the Lexicon and determining the polarity of the words.



You may not be served quickly, and the wait staff can't be called overly friendly however this is a must have New York experience. This is quality Seoul food and it's open 24/7. They have a decent selection of soju and sake. The kimchee is excellent. My personal favorite here is the dak bul gui, however the sizzling rice bowls are enticing as well and if you're into noodles try the Jap che. Don't be intimidated by the menu. Choose something that has ingredients you enjoy and don't worry about being properly satiated or satisfied. This is a great date dinning experience, not for the atmoshpere but for the adventure and food.

Figure 2(d). The example above after aggregating the word scores and determining the polarity of an entire sentence.



http://tweenis.nyc.corp.google.com:31313/data?url=http%3A//_00000000000000000000_21/47_&page_set=evaluations_arabic

فندق le kandian. فندق 4 نجوم .. الخدمه فيه كانت جيده نوعاً ما .. عيبه هو ان الغرف فيه صغيره جداً .. يوجد جهاز تلفزيون بالغرفه .. لكن التكييف كان سيء .. وجدنا في الفندق عدد من العوائل
الخليجيه ... اطلالة الفندق كانت جيده .. الطعام في الفندق كان عباره عن بوفيه مفتوح في وجبتي العشاء و الافطار .. و الاكل كان لذيذ جداً .. فأكثر ما يميز هذا الفندق هو مطعمه .. يحتوي الفندق على مسبح و
مركز للمساج و طاولة بلياردو و طاولة كيرم ...

http://tweenis.nyc.corp.google.com:31313/data?url=http%3A//_00000000000000000000_21/47_&page_set=evaluations_arabic

فندق le kandian. فندق 4 نجوم .. الخدمه فيه كانت جيده نوعاً ما .. عيبه هو ان الغرف فيه صغيره جداً .. يوجد جهاز تلفزيون بالغرفه .. لكن التكييف كان سيء .. وجدنا في الفندق عدد من العوائل
الخليجيه ... اطلالة الفندق كانت جيده .. الطعام في الفندق كان عباره عن بوفيه مفتوح في وجبتي العشاء و الافطار .. و الاكل كان لذيذ جداً .. فأكثر ما يميز هذا الفندق هو مطعمه .. يحتوي الفندق على مسبح و
مركز للمساج و طاولة بلياردو و طاولة كيرم ...

Figure 3. An Arabic example for applying the Sentiment Analysis.

Then, we used a sentence boundary detection to identify the sentences in documents, where long sentences (more than 120 characters) were discarded.

Using the Arabic Lexicon, the negation mechanism, and the sentence boundary detection, the Arabic sentiment analyzer is built. Given the scores of every word / phrase in a sentence, and the flipped scores of negated words/phrases, the scores are added together. The total score of a sentence will determine if it is positive, negative, mixed or neutral.

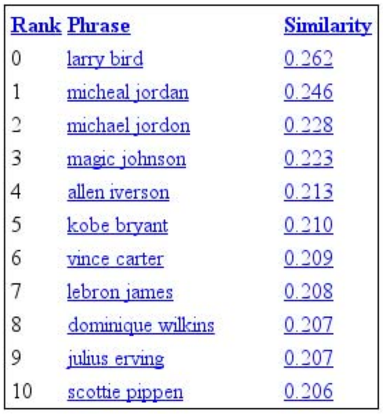
Fig. 2 shows a working example of the Sentiment Analysis process, using English text for clarification. Fig. 2(a) shows the raw input text. Fig. 2(b) shows the detected sentences (highlighted in grey) after applying the sentence boundary detection. Fig. 2(c) shows the words / phrases that were identified as having positive sentiment (highlighted in green), and those identified as having negative sentiment (highlighted in pink). Note that although the word "quickly" was originally identified as positive, it got a negative score because it was negated. Finally, Fig. 2(d) shows the aggregated polarity of the sentences. The first sentence (highlighted in yellow) is identified as mixed since it contained one positive and one negative word. The other sentences (highlighted in green) are all identified as positive since they contain more positive words than than negative ones. Fig. 3 shows the same behavior for Arabic. Note that the neutral sentences are highlighted in blue.

D. Label Propagation Capping

As mentioned above, while working on the Arabic similarity graph, which is built on rather sparse data, we discovered the following problem. Certain words in the similarity graph have very high edge weights to neighbors. Nevertheless, these words and their synonym links might be garbage. This happens somewhat infrequently in a high-quality language graph such as English, but much more often in languages with sparse data. We realized the same improvement can be used in other languages including English, to suppress this type of undesirable noise.

This problem affects building any language Lexicon since the label propagation multiplies the propagated values by edge weights, which usually correlate with the precision. Hence, very large edge weights cause excess weight to be propagated. Weird clusters like these get amplified by a chain reaction, and produce high scoring garbage words.

To solve this problem, we introduced a cap on the maximum allowed value that could be propagated at each node. For Arabic, this maximum was set to 1.0 (thus preventing any chain-reaction-like amplification of propagated values). For English, a maximum of 1.5 worked better, because the graph is more accurate and we benefit from a small degree of signal amplification. Using those caps, we were able to get a slight improvement in the English lexicon, yet a significant improvement for Arabic, especially for positive sentences. Fig. 4 shows an example of one node "michael jordan" in the English similarity graph, where only the top 10 edges are shown.



Rank	Phrase	Similarity
0	larry bird	0.262
1	micheal jordan	0.246
2	michael jordon	0.228
3	magic johnson	0.223
4	allen iverson	0.213
5	kobe bryant	0.210
6	vince carter	0.209
7	lebron james	0.208
8	dominique wilkins	0.207
9	julius erving	0.207
10	scottie pippen	0.206

Figure 4. Top 10 edges for the node “michael jordan”

IV. Performance Evaluation

First, we will evaluate the Arabic Reviews classifier and compare it to the English Reviews classifier that is known to work well. Fig. 5 shows the precision-recall curves for both Arabic and English Reviews classifiers. It shows that we were able to achieve a high-precision relatively-high-recall Arabic Reviews classifier. As mentioned before, 80% of the data are used for training and 20% are used for testing.

To evaluate the Arabic Lexicon and the Arabic Sentiment Analysis, we are going to use two measures. The *a-measure* is the mean average precision of the precision-recall curve or the area underneath the curve. It gives a rough measure for comparing curves that have reasonably good precision and recall. Note that the *a-measure* is the mean average precision over the entire x-axis length. If the recall is low (i.e., the curve only covers a fraction of the horizontal space), the *a-measure* will be proportionally low. The *f-measure* [4], a weighted measure of an optimal point on the curve. Note that we are actually using *F0.5 measure*, not *F1*, which weighs precision twice as much as recall. The *F0.5* measure is better for measuring the optimal performance on a curve.

When we evaluate the Arabic Lexicon, we sort words in the lexicon by score, and measure precision and recall consecutively at each word, starting from the highest-scoring word and going down the list. The Arabic lexicon was evaluated against 605 labeled random phrases: 500 manually labeled, and 105 auto labeled neutral phrases. Fig. 6 and 7 show the evaluation for the Lexicon and the Sentiment Analysis, respectively, also compared to the known-to-be-working-well English ones. Fig. 6(a) shows that the Arabic Lexicon has a relatively-high precision, similar to the English one, and has a low recall due to the label propagation capping. Fig. 7(a) shows that Arabic Sentiment Analysis has relatively-high precision for the sentences with positive or negative sentiments, which is more important to achieve than the mixed or neutral sentiments. That behavior is also similar to that of the English Sentiment Analysis

V. Conclusions

In this paper, we have demonstrated a system for mining Arabic business reviews from the web. The system comprises two main components: a reviews classifier that classifies any webpage whether it contains reviews or not, and a sentiment analyzer that identifies the review text itself and identifies the individual sentences that actually contain a sentiment (positive, negative, neutral or mixed) about the business getting reviewed. The system is of particular interest for languages that are of poor web content, e.g., Arabic; and can easily be extended to other alike languages.

Acknowledgment

We would like to thank Sasha Blair-Goldensohn, Kerry Hannan, Leonid Velikovich, and Jeremy Shute for their help throughout the process of developing this system. We also thank Ryan McDonald for his valuable comments regarding the paper.

References

[1] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar, “Building a Sentiment Summarizer for Local Service Reviews”, Proc. of the WWW Workshop on NLP Challenges in the Information Explosion Era, NLPIX, 2008.

[2] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Proc. of the 6th Symposium on Operating System Design and Implementation, OSDI 2004, pp. 137-150.

[3] L. Velikovich, S. Blair-Goldensohn, K. Hannan, and R. McDonald, "The Viability of Web-derived Polarity Lexicons", Proc. of the North American Chapter of the Association for Computational Linguistics, 2010.

[4] http://en.wikipedia.org/wiki/Information_retrieval#F-measure

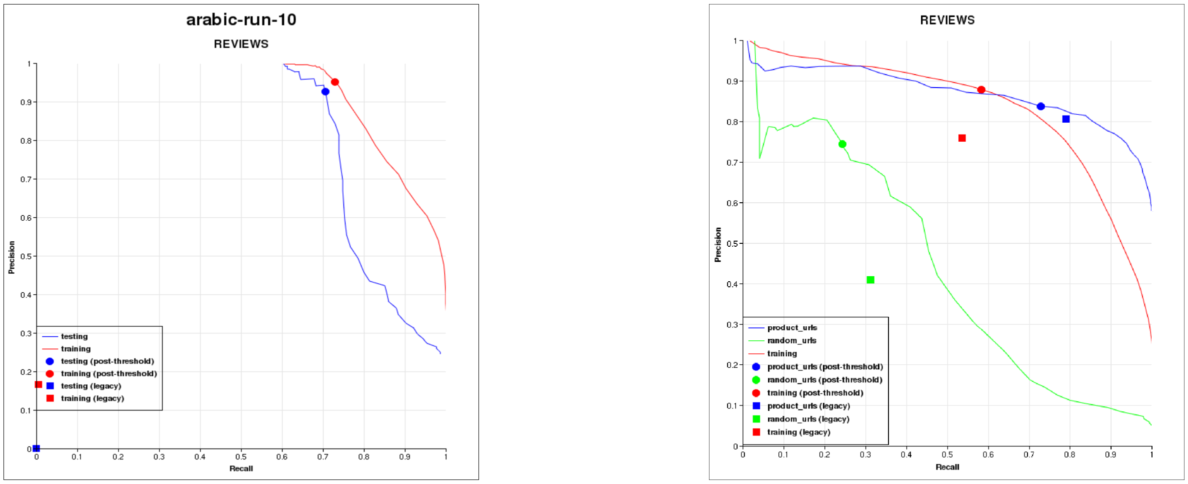


Figure 5. Performance of Reviews Classifiers: (a) Arabic (b) English

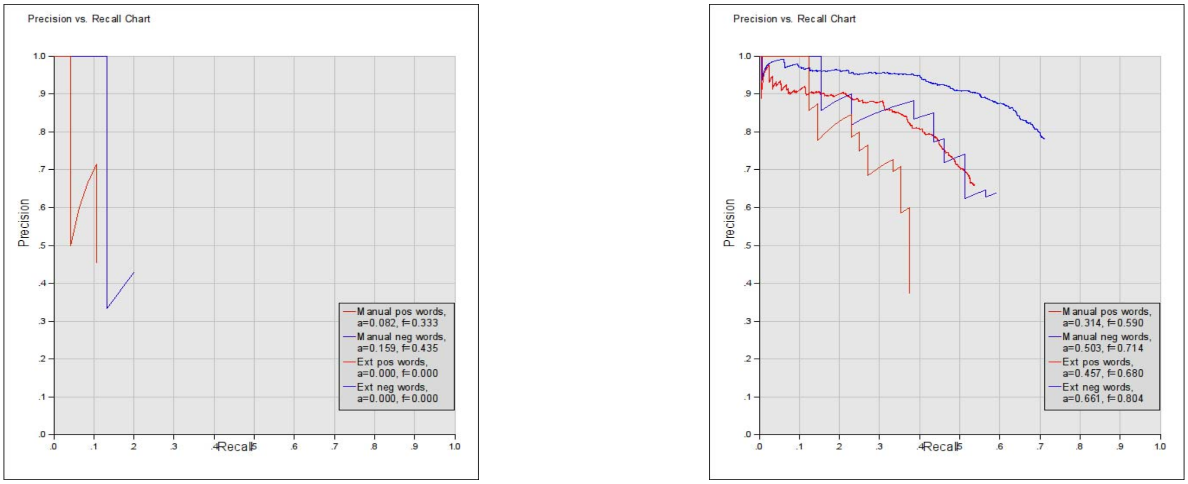


Figure 6. Performance of Lexicons: (a) Arabic (b) English

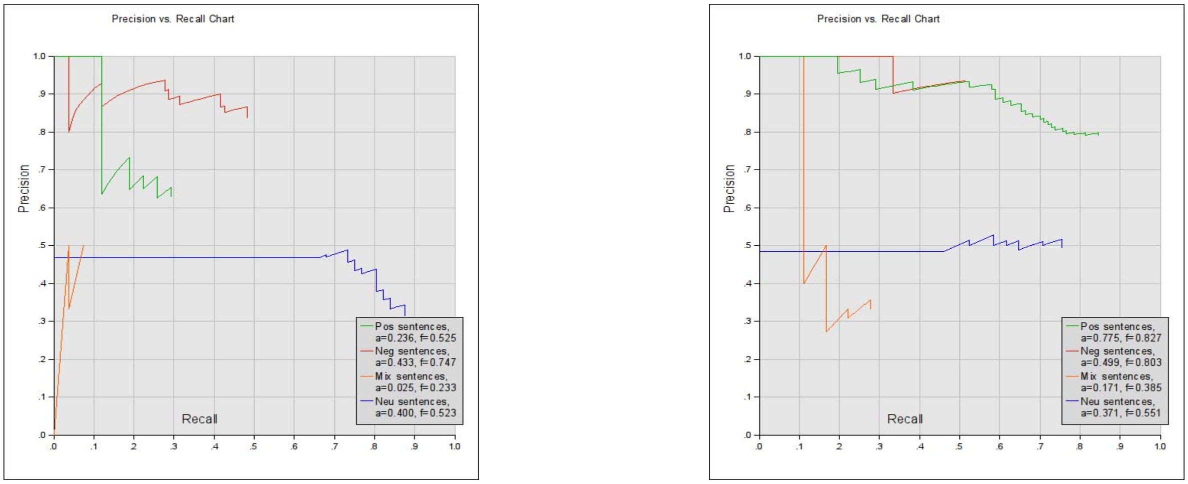


Figure 7. Performance of Sentiment Analysis: (a) Arabic (b) English