Building A Sentiment Analysis Corpus With Multifaceted Hierarchical Annotation

Muazzam Ahmed Siddiqui *maasiddiqui@kau.edu.sa*
Department of Information Systems
Faculty of Computing and Information Technology
King Abdulaziz University
Saudi Arabia

Mohamed Yehia Dahab *mdahab@kau.edu.sa*
Department of Computer Science
Faculty of Computing and Information Technology
King Abdulaziz University
Saudi Arabia

Omar Abdullah Batarfi *obatarfi@kau.edu.sa*
Department of Information Technology
Faculty of Computing and Information Technology
King Abdulaziz University
Saudi Arabia

Abstract

A corpus is a collection of documents. An annotated corpus consists of documents or entities annotated with some task related labels such as part of speech tags, sentiment etc. While it is customary to annotate a document for a specific task, it is also possible to annotate it for multiple tasks, resulting in a multifaceted annotation scheme. These annotations can be organized in a hierarchical fashion, if such a scheme naturally occurred in the data, resulting in a hierarchical text categorization problem. We developed a multifaceted, multilingual corpus for hierarchical sentiment analysis. The different facets include hierarchical nominal sentiment labels, a numerical sentiment score, language, and the dialect. Our corpus consists of 191K reviews of hotels in Saudi Arabia. The reviews are divided into eleven different categories. Within each category, the reviews are further divided into two positive and negative categories. The corpus contains 1.8 million tokens. Reviews are mostly written in Arabic and English but there are instances of other languages too.

Keywords: Multifaceted Text Categorization, Hierarchical Text Categorization, Sentiment Analysis, Corpus Linguistics, Arabic Natural Language Processing, Text Mining.

1. INTRODUCTION

Sentiment analysis refers to the identification of sentiment associated with text. Sentiment polarity classification is a subtask that limits the analysis to the identification of the polarity, as in positive or negative of the text. A fine grained annotation scheme can also be employed where the sentiment can be identified as belonging to the set of six different emotions. Reviews of hotels, movies and other objects take a slightly different approach, where the sentiment is expressed through a rating scale of least favorable to most favorable, a star rating, a numerical score or a combination of these. Hierarchical classification [1] refers to the classification scheme where the labels naturally form a hierarchy. Web directories such as DMOZ [2] and Internet Public Library [3] and Wikipedia are two examples of such hierarchies. The annotation schemes are generally task oriented, therefore the above mentioned scheme would be considered as hierarchical sentiment analysis. On the other hand, a multifaceted approach annotate the same text with different labels associated with different tasks. Such an approach is referred to as multifaceted

text categorization [4]. A document can be categorized based upon the sentiment it bears, but it can also be categorized based upon the language or topic etc.

In this paper, we present the ongoing effort to develop a multifaceted, multilingual corpus for hierarchical sentiment analysis. The corpus consists of more than 191K reviews of hotels in Saudi Arabia in mainly two different languages, Arabic and English. The corpus contains 1.8M tokens. Each document (review) in our corpus is annotated along the following three facets.

1. Sentiment label: A hierarchical nominal sentiment label with two levels. The first level assign one of the eleven rating labels ranging from Exceptional to Very poor, while the second level classify the review as being positive or negative within the first level label.
2. Sentiment score: A numerical score from 1 to 10 representing the sentiment.
3. Language/dialect: A hierarchical language label with two levels. The first level reflect the main language of the review, and in case of Arabic, the dialect of Arabic in the second level.

2. RELATED WORK

A corpus is a valuable resource for linguistics research. It is used to test different hypothesis about language use, to test and generate linguistic rules, and to build predictive models. The statistical natural language processing approach relies on the presence of a corpus to induce a language model using statistical, pattern recognition and machine learning methods [5]. In this section, we will cover notable works to create corpora for dialect modelling and sentiment analysis, two facets of annotation supported by our corpus.

There are many Arabic dialects in the Arab world. These dialects vary from region to region and maybe from city to city. Arabic dialects differ, as modern standard Arabic (MSA), on all levels of linguistic representation, phonology, morphology and lexicon to syntax. The extreme differences are on phonological and morphological levels [6].

For multidialectal Arabic corpora, guidelines for the construction of large corpora of multidialectal Arabic resources are provided in [6] and [7]. There are several multidialectal Arabic corpora such as [8], [9], [6] and [10] [11]. All of them are manually annotated for at most five MSA dialects. Recently research has focused on corpus of classic Arabic [12] too.

Corpus subjectivity and sentiment analysis (SSA) sources maybe:

- Reviews product such as movie, and music reviews
- Web discourse such as web forums and blogs
- News articles such as online news articles and web pages
- Social media websites such as Twitter, Facebook, YouTube

News corpora are manually labeled for SSA at the word and phrase levels [13], [7] and [14]. While [15] and [16] described labeling a collection of documents from Arabic Web forums. Besides corpus, another valuable resource in Arabic sentiment analysis is a sentiment lexicon. Efforts to build Arabic sentiment lexicon are described in [17] and [18].

3. CORPUS CREATION

Our corpus was created by crawling a popular hotel review website[1]. The corpus contains reviews of more than 650 hotels in Saudi Arabia. The website uses Ajax [19] to dynamically display the review contents in one part of the page while the rest of the page with the overall score, review summary and other hotel details remain unchanged. Each page displays ten reviews and clicking on the Next Page hyperlink loads the next ten reviews in the review area,

[1] Due to copyright and privacy concerns, we will not disclose the name of the website. The corpus will be used for strictly research purposes

while keeping the rest of the page unchanged. We built a simple web crawler/scrapper that given a seed URL scrape all or a subset of the pages from a website given some criteria. We were able to identify a static URL from the Next Page hyperlink which can be used to display the review texts without any formatting. By changing a parameter value in the URL we were able to get all the reviews of a hotel in one HTML page. The URL also includes the name of the hotel. The hotel names were manually identified and a list of URLs to crawl was prepared in advance. The total number of reviews for a hotel was also manually identified by going to the webpage of each hotel on the review website. The list of URLs was given as an input to the crawler that went to each page, downloaded the HTML, extract the text and saved it locally as a UTF-8 text document. We used simple lexical patterns to parse the HTML page and extract the reviews and the associated annotation information. The lexical patterns made use of the class name of the HTML elements defined in the CSS style sheets.

A review on the website consist of the following items:

1. A numerical score from 1 to 10
2. An optional title of the text review
3. A nominal rating label chosen from the one of the eleven available categories
4. An optional text review indicating the positive aspects
5. An optional text review indicating the negative aspects

No	Category Arabic	Category English
1	استثنائي	Exceptional
2	ممتاز	Excellent
3	رائع	Wonderful
4	جيد جدا	Very good
5	جيد	Good
6	مرضي	Pleasant
7	حسن	Fair
8	مقبول	Okay
9	مخيب للأمل	Disappointing
10	ضعيف	Poor
11	ضعيف جدا	Very poor

TABLE 1: Review categories in Arabic and English.

The reviewer can choose the rating label from one of the eleven available categories as described in Table 1, or provide his/her own short and concise title of the review. In case of the latter, the rating label is determined from the numerical score. The label is not displayed on the page, but can be extracted from the source HTML. The actual body of the review is optional but if the reviewer chooses to provide one, the positive and negative aspects are written separately. It is clear from the above mentioned scheme that the review is already annotated by the reviewer. In the rest of the paper we will refer to the overall review containing all or a subset of the above mentioned five items as review, while the option positive text will be referred to as positive review and the option negative text will be referred to as negative review. If N is the set of reviews, p is a positive review and n is a negative review, then it should be noted that $n \epsilon N$, $p \epsilon N$ and $\Sigma p + \Sigma n \neq \Sigma N$.

4. CORPUS ANALYSIS

The corpus was analyzed to compute the basic descriptive statistics. This section will provide statistics including the document and token level statistics from the corpus.

4.1 Corpus Statistics

We made a distinction between a review and a document in our corpus. As described previously, a review may consists of a positive review and a negative review along with the nominal label and the numerical score. To build the corpus, we considered each positive or negative review as a document. The total number of documents and tokens in the corpus are given by Table 2. A token is each individual word. Although a number of Arabic tokenizers are available, for the purpose of statistics reporting we employed a simple space based tokenization. For later text processing we are planning to use MADAMIRA [20] for a full-fledged morphological analysis.

Type	Value
No of documents	191,011
No of tokens	1,830,191

TABLE 2: Number of documents and tokens in the Corpus.

4.2 Review Statistic

The number of hotels, the number of reviews and other related statistics are described in Table 3. The number of reviews per hotels followed a power law distribution as displayed by Figure 1. This indicates that there were few hotels that received a large number of reviews, while a large number of hotels received few reviews.

Type	Value
No of hotels	658
No of reviews	176,884
No of reviews containing both positive and negative reviews	80,506
No of reviews with no positive or negative review	66,837
No of reviews with at least one positive or negative review	29,541
No of positive reviews	95,884
No of negative reviews	95,127
Total no of positive and negative reviews	191,011

TABLE 3: Number of positive, negative and total reviews.

It can be noted from the descriptive statistics provided in Table 4, that the median number of reviews for a hotel is 65.5. The mean is much higher than the median indicating a positive skew in the data. There were only 46 hotels with more than 1000 reviews but because of the presence of this, the mean is much higher than the median.

Statistic	Value
Mean	268.8
Median	65.5
First Quartile	17
Third Quartile	238.2
Maximum	5390
No of hotels with more than 1000 reviews	46

TABLE 4: Descriptive statistics of the number of reviews per hotel.

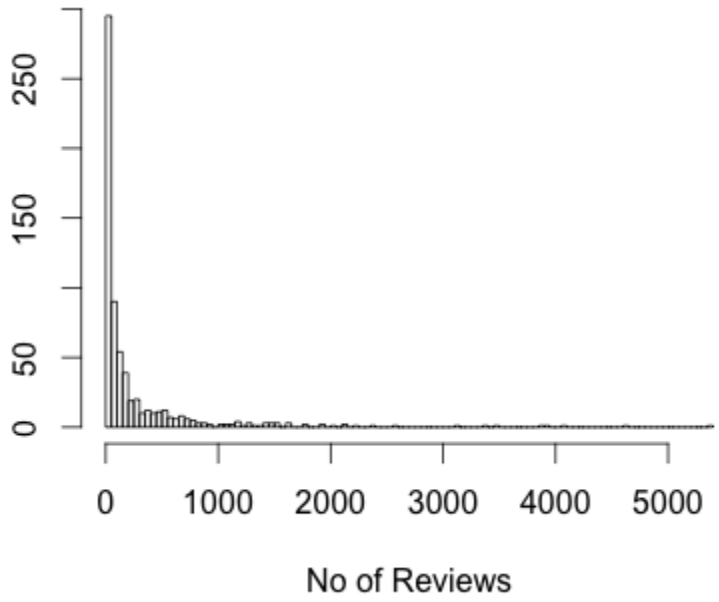

FIGURE 1: Distribution of the number of reviews per hotel.

4.3 Annotation Statistics

This section describes the statistics related to the multifaceted, multilingual, hierarchical categorization scheme of annotation. Figure 2 and Figure 3 display the number of reviews and the number of positive and negative reviews for each category. It is interesting to note that in both the figures, most of the reviewers wrote favorable reviews and exceptional and good constitutes the majority categories. Figure 3 displays a comparison of the number of positive and negative reviews in each category. It can be noted that for the good category, the number of positive and the negative reviews is almost the same. Categories with higher rating have more positive reviews than negative and categories with lower rating have more negative than positive reviews. This is intuitive as a person writing a favorable review will not find many negative aspects and is more likely to write the positive review only and same is true for an unfavorable review. This is more pronounced in Figure 4, where the absolute value of the difference between positive and negative reviews, normalized by the sum of positive and negative reviews is plotted. Let p_i be the number of positive review and n_i be the number of negative reviews in category i, then the absolute normalized difference d_i for category i is given by the equation 1.

$$d_i = \frac{|p_i - n_i|}{p_i + n_i} \tag{1}$$

The mean absolute difference $\bar{d}$ between number of positive and negative reviews for each category is 1480.36, computed using equation 2, where C is the number of categories.

$$\bar{d} = \frac{\sum_i^C |p_i - n_i|}{C} \tag{2}$$

The average score for each category is given by Figure 5.

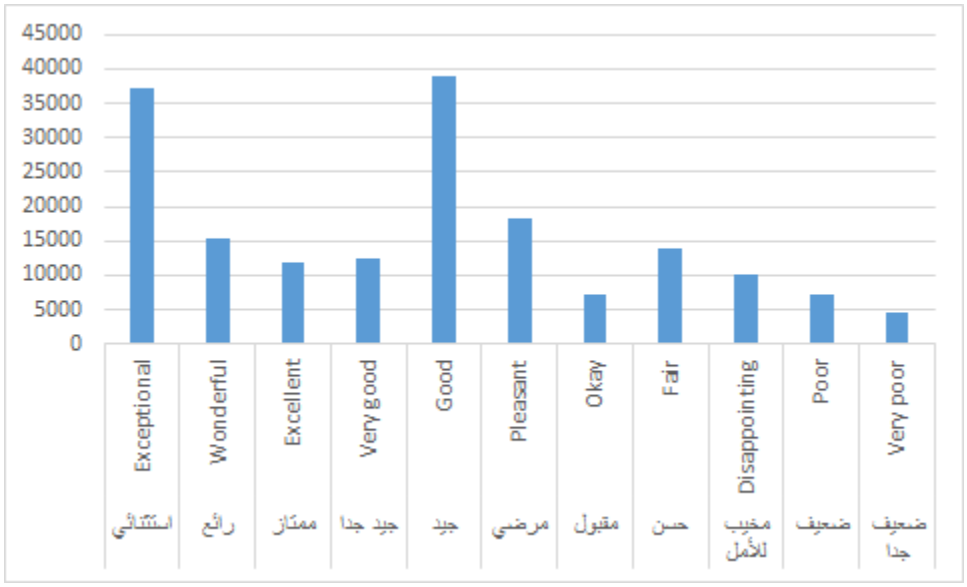

FIGURE 2: Number of reviews in each category.

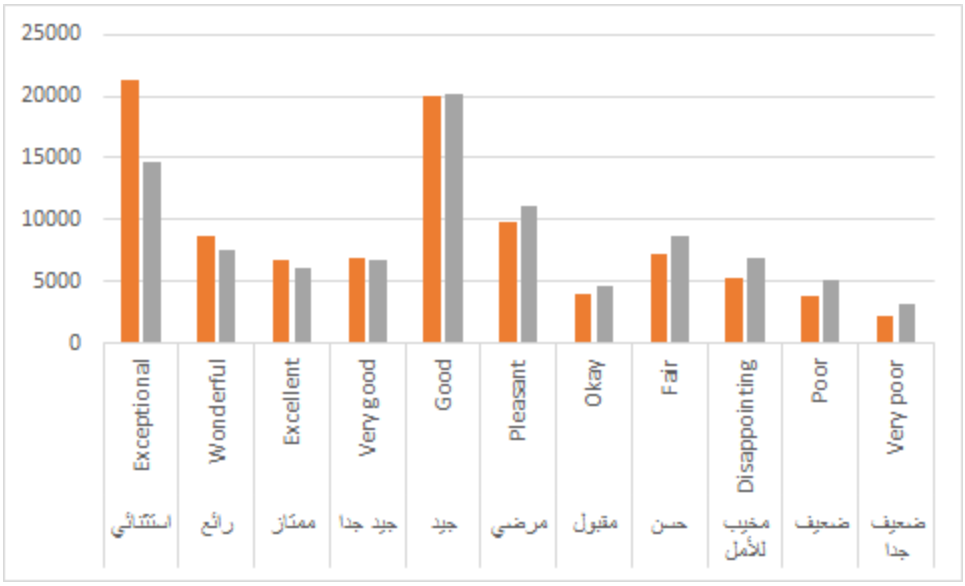

FIGURE 3: Number of positive (orange) and negative (grey) reviews for each category.

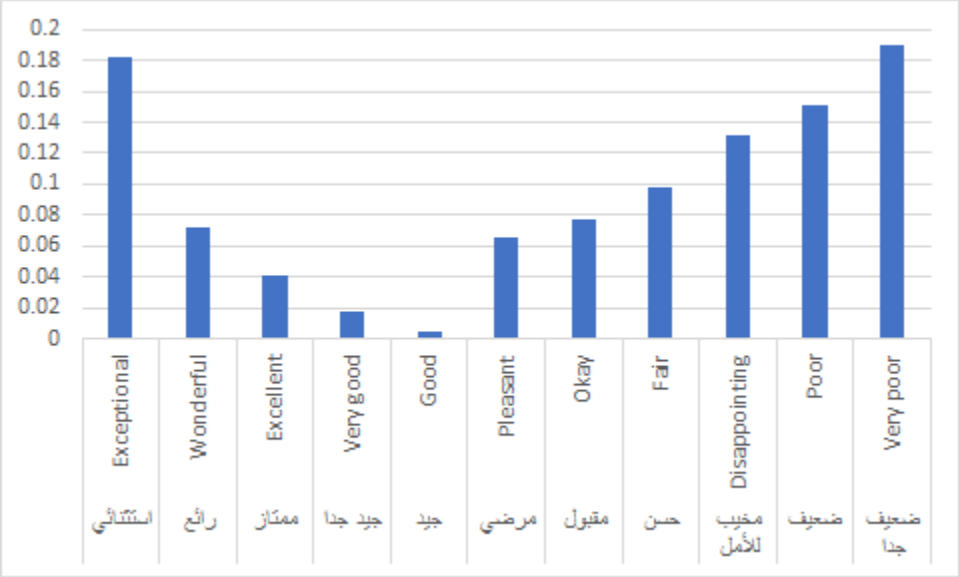

FIGURE 4: Normalized absolute difference between number of positive and negative reviews for each category.

The second facet of our annotation is the language. We used a combined automatic and manual method to annotate each review with its language. In the first step we automatically identified the language of the review using the Unicode character value of the first character of each review. This is a naïve method as it is based upon the assumption that the entire review was written in one language. A manual validation was carried out in the next step, to check the correctness of the assumption. It was revealed that the assumption held with two exceptions. One, where the review was mainly written in Arabic but few English words were interspersed between Arabic words. And two, where the reviewer enumerated the review and used the Arabic numerals [21] for enumeration, instead of the Eastern Arabic numerals [22], while the review was actually written in Arabic. Please note that the Arabic numerals are the most common representation of the digits used in English and to distinguish from the digits used in Arabic, the latter is referred to as Eastern Arabic numerals. For the first case, we annotated the review with the Arabic, instead of creating a new annotation category called mixed. The second case was manually fixed. Figure 6 displays the number of positive and negative reviews in each language. Besides the main language, our goal was also to annotate the Arabic reviews with their dialect. We hypothesize that we can infer the dialect using the home country of the reviewer, a piece of information that we have already extracted from each review. Figure 7 displays the number of positive and negative Arabic reviews from reviewers belonging to Arab countries. It can be seen from Figure 7, that majority of the reviewers were from Saudi Arabia. To get a clearer picture of reviewers from other Arabic countries, Figure 8, displays the number of positive and negative reviews from reviewers belonging to Arab countries except Saudi Arabia. According to [23], there are about 2.4 million expatriate Arabs living in Saudi Arabia. We can safely assume that they all speak their own Arabic dialects, so it is not possible to infer the dialect from the country, in the case of Saudi Arabia, as the only meta information available to us is the country of the reviewer. The dialects of Arabic spoken in different countries are presented in Table 5.

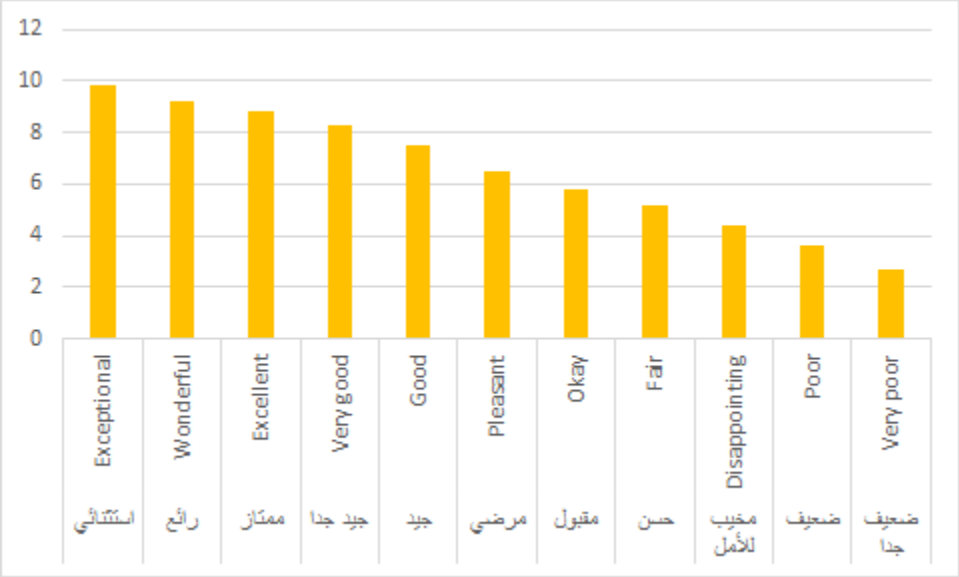

FIGURE 5: Average score for each category.

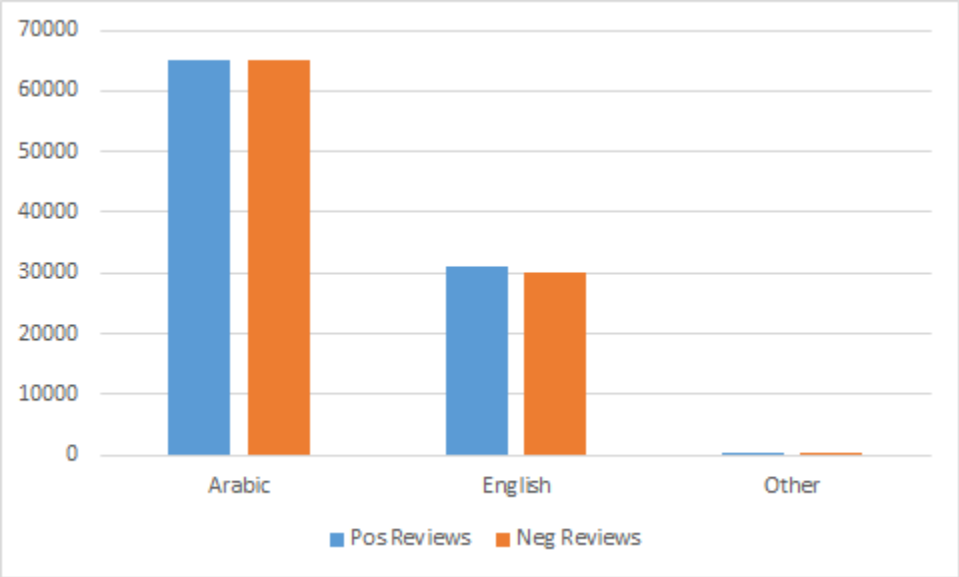

FIGURE 6: Number of positive and negative reviews in each language.

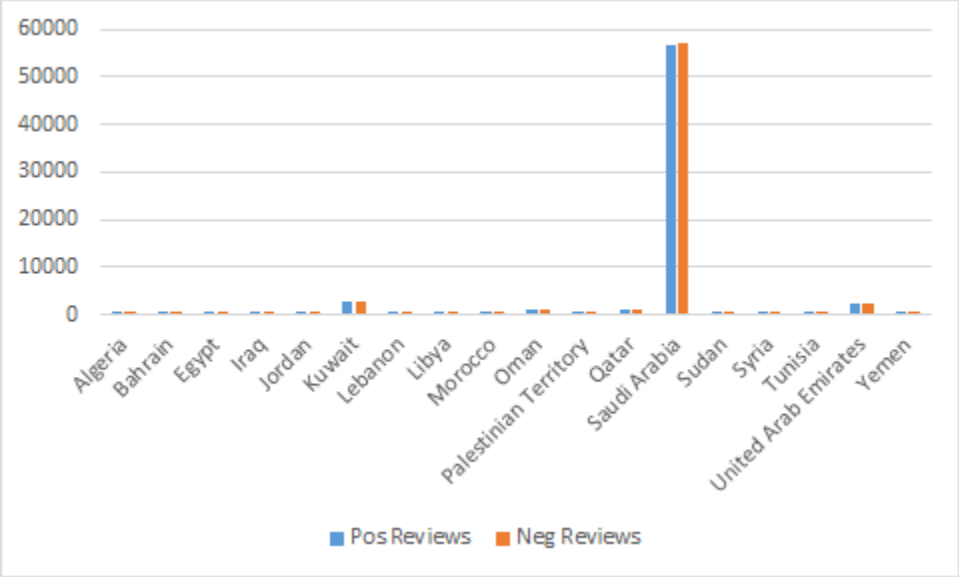

FIGURE 7: Number of positive and negative reviews in Arabic from reviewers belonging to Arab countries.

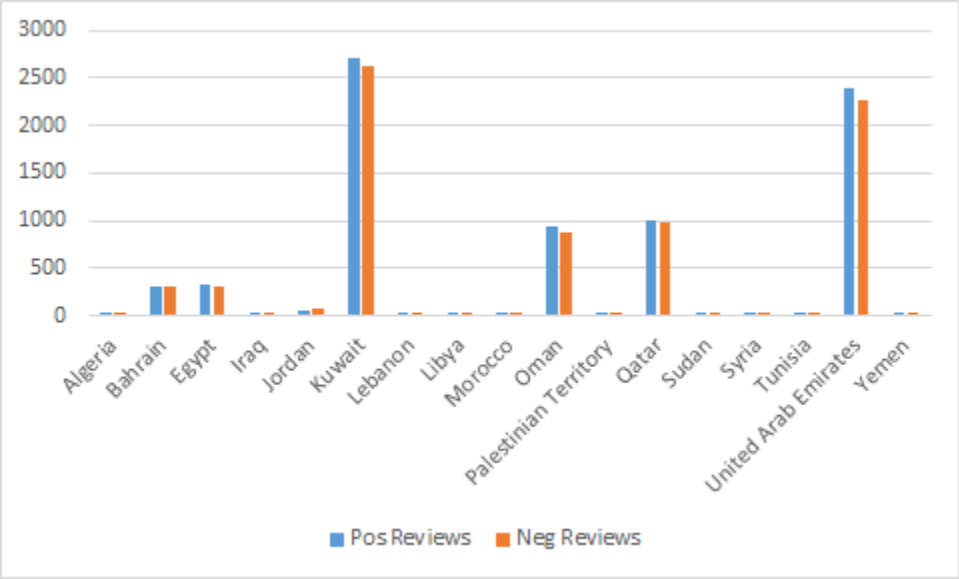

FIGURE 8: Number of positive and negative reviews in Arabic from reviewers belonging to Arab countries except Saudi Arabia.

No	Country	Dialect	No	Country	Dialect
1	Oman	Gulf	10	Palestinian Territory	Levantine
2	Saudi Arabia	Hijazi, Najdi	11	United Arab Emirates	Gulf
3	Yemen	Yemeni	12	Iraq	Iraqi
4	Egypt	Egyptian	13	Sudan	Sudanese
5	Kuwait	Gulf	14	Libya	Maghreb
6	Jordan	Levantine	15	Qatar	Gulf
7	Morocco	Maghreb	16	Bahrain	Gulf
8	Algeria	Maghreb	17	Lebanon	Levantine
9	Syria	Levantine	18	Tunisia	Maghreb

TABLE 5: Dialects spoken in different Arab countries.

Except for Saudi Arabia, we annotated the Arabic positive and negative reviews with dialects using the information from Table 5. The number of positive and negative reviews in our corpus in each Arabic dialect is displayed in Figure 9.

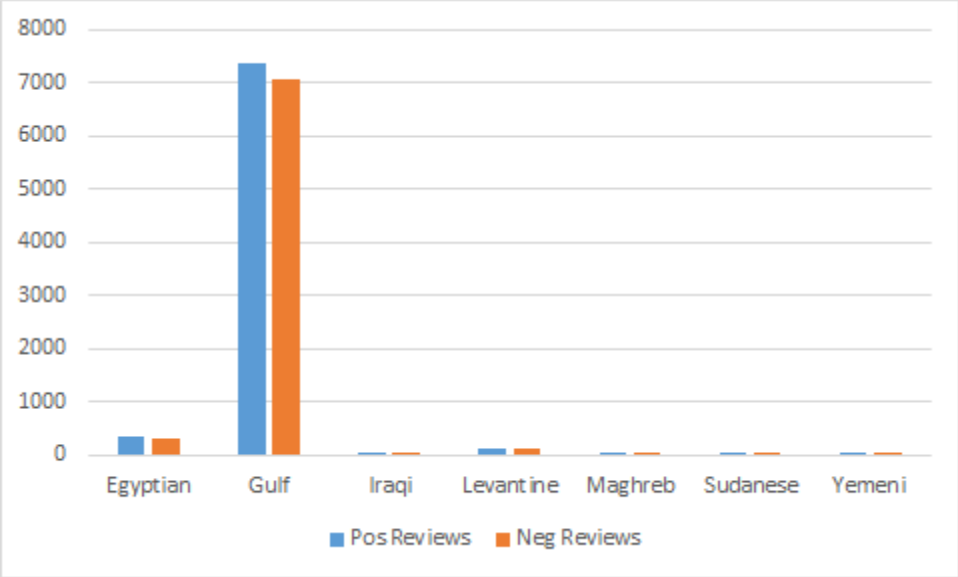

FIGURE 9: Number of positive and negative reviews in each Arabic dialect.

4.4 Token Statistics

The corpus contains more than 1.8M tokens in 191K documents (positive and negative reviews). We will define the length of a document as the number of tokens present in that document. Both, the positive and negative reviews were usually short with mean length of 7.4 tokens for the positive documents and 11.7 for negative documents. Other basic statistical descriptors for positive and negative reviews can be found in and Table 6. It is evident from the table that negative reviews are slightly longer than the positive reviews. This indicates that when people write a bad review they are more elaborate than when they are writing a good review. There are few long reviews in the corpus too which is evident from the maximum length and the fact that mean is large than the median. The distribution of length of positive and negative reviews in Figure 10 and Figure 11 clearly display the right skew, as a result of the presence of outliers.

	Originals		Without Outliers	
Statistic	**Positive Reviews**	**Negative Reviews**	**Positive Reviews**	**Negative Reviews**
Mean	7.44	11.74	5.36	8.34
Median	5	7	4	6
Standard Deviation	9.43	15.06	4.01	6.59
First Quartile	3	4	2	3
Third Quartile	9	14	7	12
Minimum	1	1	1	1
Maximum	316	360	18	29

TABLE 6: Basic statistical description of the lenght of positive and negative reviews, before and after removing outliers.

We report the statistics without outliers also in Table 6, to get a clearer picture of review lengths. The outliers were identified using interquartile range. Let *Q1* be the first quartile and *Q3* be the third quartile, a document is considered to be an outlier, if the length fell outside the following *range [Q1 – k * (Q3 – Q1), Q3 + k * (Q3 – Q1)]*. We used *k*=1.5 in our calculation. Based upon this, the statistics without outliers are given in the right two columns of Table 6. The distribution of the lengths of positive and negative reviews without outliers is given by Figure 11 and Figure 12. Even after removing the longer reviews, it is evident the lengths still do not follow normal distribution. About 64% of the positive reviews are shorter than the mean length of the positive reviews. Similarly, 63% of the negative reviews are shorter than the mean length of the negative review.

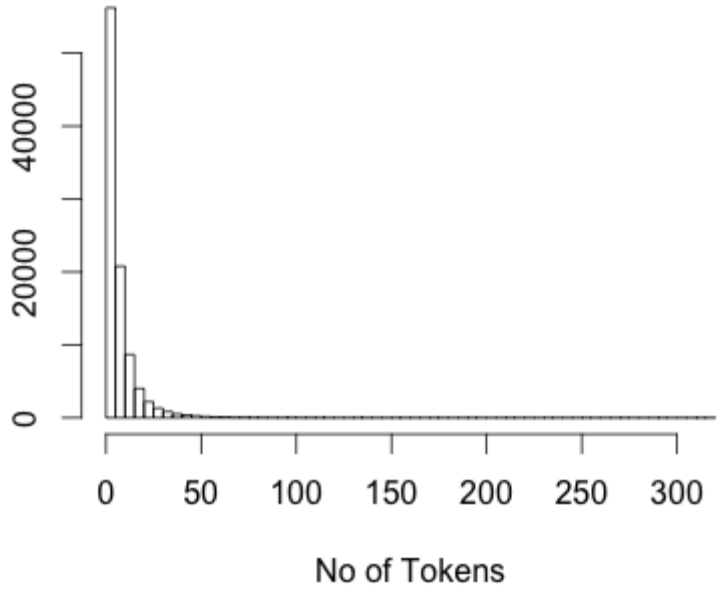

FIGURE 10: Distribution of the length of positive reviews.

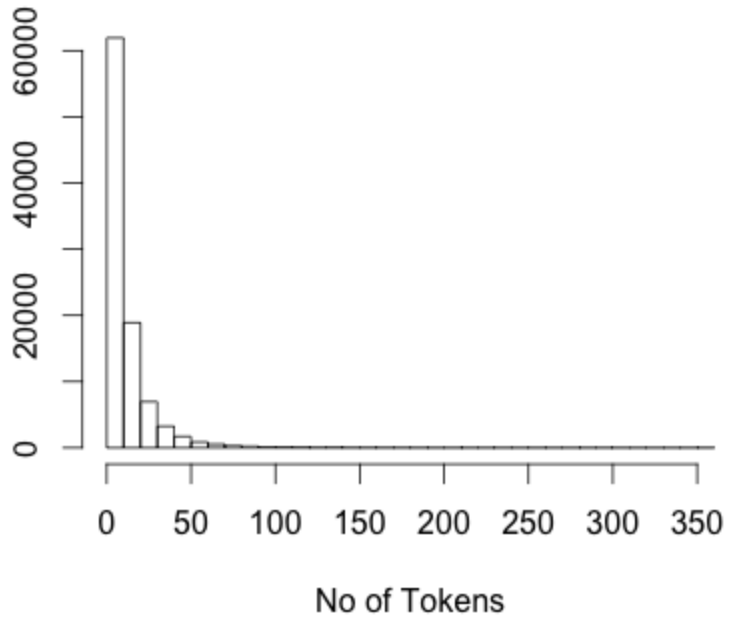

FIGURE 11: Distribution of the length of negative reviews.

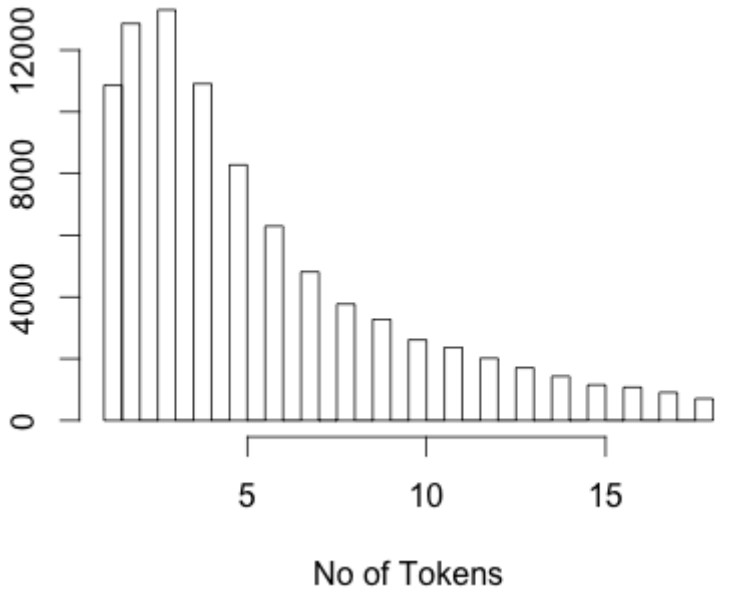

FIGURE 12: Distribution of the length of positive reviews with no outliers.

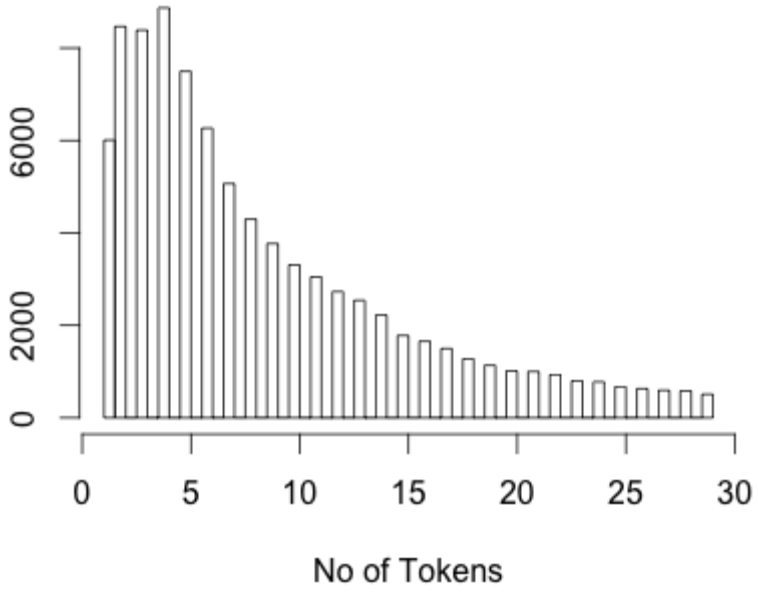

FIGURE 13: Distribution of the length of negative reviews with no outliers.

Another important dimension to plot the document length against is the main category of the review. Figure 14 display the plot of normalized lengths of positive and negative reviews against the 11 categories. The lengths were normalized by dividing the number of tokens by the number of reviews in each category. A decreasing trend while going from the Exceptional to the very poor category can be observed for the length of positive reviews, while the opposite is true for negative reviews.

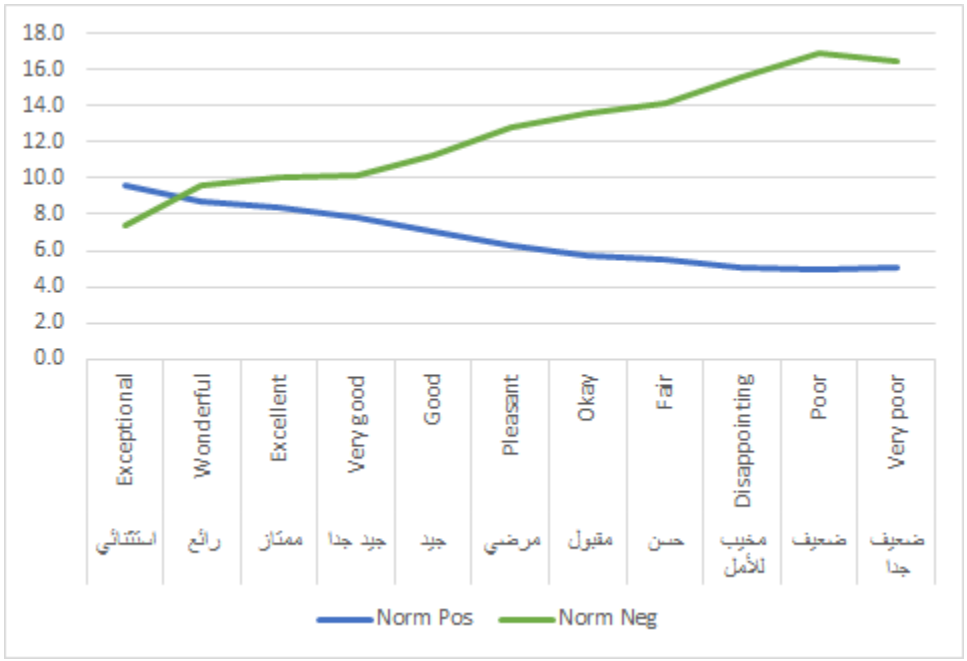

FIGURE 14: Normalized length of positive and negative reviews for each category.

5. CONCLUSION

This paper discusses the development effort and the statistics for a sentiment analysis corpus. The corpus has a multifaceted annotation including hierarchical sentiment polarity, sentiment score, language and dialect. The corpus consists of more than 191K hotel reviews written in colloquial Arabic and English. Each review is annotated as positive or negative at the lower level, while at the higher level of hierarchy, one of the 11 categories are used to annotate the review ranging from Exception to Very poor on a rating scale. We presented different statistics including the number of reviews per category, length of reviews, number of reviews per language and per dialect etc. Building this corpus is part of our ongoing research to study the effect of dialect on Arabic sentiment analysis. The corpus will serve as a gold standard for dialect modeling and building sentiment classifiers which will incorporate the effect of dialect and word collocations into account. Barring any legal issues, we are planning to release the corpus for other researchers interested in investigating sentiment analysis in Arabic.

6. ACKNOWLEDGEMENTS

This work was supported by a King Abdulaziz City of Science and Technology (KACST) funding (Grant No. 12-INF2751-03). We thank KACST for their financial support

7. REFERENCES

[1] A. D. Gordon, "A Review of Hierarchical Classification," Journal of the Royal Statistical Society. Series A (General), vol. 150, no. 2, pp. 119-137, 1987.

[2] "DMOZ," Open Directory Project, [Online]. Available: http://www.dmoz.org/. [Accessed 20 4 2015].

[3] "Internet Public Library," ipl2, [Online]. Available: http://www.ipl.org/. [Accessed 20 4 2015].

[4] W. Dakka, P. Ipeirotis and K. Wood, "Automatic construction of multifaceted browsing interfaces," in Proceedings of the 14th ACM international conference on Information and knowledge management , 2005.

[5] C. Manning and H. Schutze, Foundations of Statistical Natural Language Processing, MIT Pres, 1999.

[6] M. Diab, N. Habash, O. Rambow, M. Altantawy and Y. Benajiba, COLABA: Arabic dialect annotation and processing., LREC Workshop on Semitic Language Processing, 2010.

[7] H. Elfardy and M. Diab, "Simplified guidelines for the creation of Large Scale Dialectal Arabic Annotations.," in Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), 2012.

[8] R. Al-Sabbagh and R. Girju, "YADAC: Yet another Dialectal Arabic Corpus," in Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), 2012.

[9] R. Cotterell and C. Callison-Burch, "A multi-dialect, multi-genre corpus of informal written Arabic," in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014.

[10] O. Zaidan and C. Callison-Burch, "The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies:

short papers - Volume 2 (HLT '11), Vol. 2, 2011.

[11] H. Elfardy, M. Al-Badrashiny and M. Diab, "Code Switch Point Detection in Arabic," in Natural Language Processing and Information Systems, Springer, 2013, pp. 412-416.

[12] M. Alrabiah , A. Al-Salman, A. Al-Salman and E. Atwell, "An Empirical Study On The Holy Quran Based On A Large Classical Arabic Corpus," International Journal of Computational Linguistics (IJCL), vol. 5, no. 1, pp. 1-13, 2014.

[13] M. Abdul-Mageed, M. Diab and S. Kubler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media.," Computer Speech & Language, vol. 28, no. 1, pp. 20-37, 2014.

[14] J. Wiebe, T. Wilson and C. Cardie, "Annotating Expressions of Opinions and Emotions in Language," Language Resources and Evaluation, vol. 39, no. 2-3, pp. 165-210, 2005.

[15] A. Abbasi, C. Hsinchun and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums.," ACM Transactions on Information Systems, vol. 26, no. 2, 2008.

[16] M. Abdul-Mageed and M. Diab, "Subjectivity and sentiment annotation of modern standard arabic newswire.," in Proceedings of the 5th Linguistic Annotation Workshop (LAW V '11), 2011.

[17] F. Mahyoub, M. Siddiqui and M. Dahab, "Building an Arabic Sentiment Lexicon Using Semi-supervised Learning," Journal of King Saud University - Computer and Information Sciences, vol. 26, no. 4, pp. 417-424, 2014.

[18] S. Alhazmi, W. Black and J. McNaught, "Arabic SentiWordNet in Relation to SentiWordNet 3.0," International Journal of Computational Linguistics (IJCL), vol. 4, no. 1, pp. 1-11, 2013.

[19] J. Garrett, "Ajax: A New Approach to Web Applications," 18 2 2005. [Online]. Available: http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications/. [Accessed 20 4 2015].

[20] A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow and R. Roth, "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014.

[21] "Arabic numerals," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Arabic_numerals. [Accessed 20 4 2015].

[22] "Eastern Arabic numerals," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Eastern_Arabic_numerals. [Accessed 20 4 2015].

[23] A. Kapiszewski, Arab Vs Asian Migrant Workers in the GCC countries, 2006.