Birzeit University

Electrical and Computer Systems engineering Department

SP.TOP NATURAL LANGUAGE PROCESSING (NLP) AND INFORMATION RETRIEVAL ENCS539

Course Term Project

Arabic Sentiment Analysis System with NLP Operations

Student name: Yazeed Obaid Student No.: 1130036

Student name: Abdallatif Sulaiman Student No.: 1131090

Student name: Hussein Dahir Student No.: 1131138

Instructor name: Dr. Adnan Yahya

Section No.: 1

Date: 23/01/2018

Contents

Abstract:

Processing natural text and analyzing the meaning behind that text is now one of the important and challenging tasks in Natural Language Processing (NLP) community. This is due to the importance of this analysis for many parities, from researchers to commercial companies. This analysis will help in building better and more accurate systems. One of NLP text understanding and analysis is the task of Sentiment Analysis. Sentiment analysis is the task of identifying the intention or the emotional state of the user utterance. In this report we present our work on an Arabic sentiment analysis system that uses the term-frequency and inverse-term frequency as a feature to a Support Vector Machine (SVM) classifier. The system was tested on six datasets.

Introduction:

Sentiment Analysis refers to the use of natural language processing and text analysis to systematically identify, extract, and study affective states and subjective information. Sentiment analysis is widely applied to reactions and thoughts of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine [1].

Generally speaking, sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation, affective state (that is to say, the emotional state of the author or speaker), or the intended emotional communication (that is to say, the emotional effect intended by the author or interlocutor) [1].

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral [1].

Sentiment analysis for a language is usually dependent on manually or semi-automatically constructed lexicons, found in dictionaries or corpora. The availability of these resources enables the creation of rule-based sentiment analysis or the construction of training data for classification tasks [2]. For Arabic language, it didn't gain much attention in NLP in general and Sentiment Analysis task in specific since it lacks of resources and data-sets. But today's social media and open internet, the data is available and Arabic is start to gain attention. In addition, Arabic as a complex morphology due to its derivational and inflectional properties which makes it difficult to deal with.

Twitter is one of the biggest microblogging services on the internet. Microblogs are short text messages that people use to share all kinds of information with the world. On Twitter, these microblogs are called "tweets", and over 400 million of them are posted every day. They can contain news, announcements, personal affairs, jokes, opinions and more [3].

Related Work:

There are many studies have been done in opinion mining field. Most of these studies have been done in English language context, and a little in Arabic language context. We next present some studies of Arabic language context. Existing approaches to sentiment analysis can be grouped into three main categories: knowledge-based techniques, statistical methods, and hybrid approaches.

Knowledge-based techniques classify text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid, and bored. Some knowledge bases not only list obvious affect words, but also assign arbitrary words a probable "affinity" to particular emotions [4] [5].

Statistical methods leverage on elements from machine learning such as latent semantic analysis, support vector machines, "bag of words" and Semantic Orientation — Pointwise Mutual Information. The most widely used methods (by far) appear to be based on Support Vector Machines (SVM), Naive Bayes (NB), and K-Nearest Neighbors (KNN). Hybrid approaches leverage on both machine learning and elements from knowledge representation such as ontologies and semantic networks in order to detect semantics that are expressed in a subtle manner [4] [5]. The Arabic sentiment analysis systems developed have reached an 80% accuracy approximately in all of the approaches used. In the next section we describe our approach to sentiment analysis in Arabic.

Approach:

Every NLP and machine learning system can be divided into three main stages: **pre-processing stage,** in this stage the data-set is pro-processed by a set of operations to put it in a form that is easy for the next two stages to work on. The second stage is the **feature extraction stage**, in which a set of features are defined and extracted from the data-set. Those features will be used in the last stage. The **Training a machine learning model stage**. In this stage, the features extracted from data-set are used as an input to the machine learning algorithm to train its parameter. Our system follows this architecture. In the next sub-sections, we describe our system from this point of view.

Pre-processing:

A pre-processing operations are needed to process the data-set and reduce the variations and un-necessary components in the data-set. The importance of this stage comes from its effect on the machine learning and the learning process. As the data-set has a clear and un-ambiguous structure, this will help and increase the accuracy of the learning process of the machine learning model. In our system, we have used a set of pre-processing tasks, we explain them below.

Normalization, normalizing the data-set means to remove un-necessary symbols and letters from the data-set examples. Our normalizer is based on **pyArabic** Python package. We used pyArabic **strip tatweel**, **strip tashkeel**, **strip harakat** and **normalize Hamza** components. The name of each component explains its task.

Stop word removal, stop words are usually referring to the most common words in a language. These words are removed from data-set since them are the majority of the words and will reduce the presence of unique and important words. We used Python's **NLTK** pre-defined list of Arabic stop words.

Stemming, to stem a word means to reduce it to its root. This will reduce the variations between words and reduce them to a unified word. We used Python's **NLTK ISRI Arabic stemmer.**

These operations are important since we use data-sets that contains customer reviews, and these reviews often have symbols that are un-necessary and if they removed don't affect the meaning of the review.

Feature Extraction:

We used the **Term-Frequency Inverse-Term-Frequency (TF-IDF)** as a feature in our system. The TFIDF, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The tf-idf value increases proportionally to the number of times a word appears in the document, but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Nowadays, tf-idf is one of the most popular term-weighting schemes. To calculate the TF-IDF features, we used **scklearn** Python package **CountVectorizer** and **TfidfTransformer** components.

Machine Learning:

After preparing the data-set and extracting the features from it, we now used it in training a machine learning model. We choose to use the Support Vector Machine (SVM) and Naïve Bias (NB) algorithms in the learning process. We choose them since they are simple and are proven to give high accuracy results.

A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outlier's detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Abstractly, naive Bayes is a conditional probability model.

In the next section we present our results from applying the above approach to the data-sets we used.

Experiments and Results:

We used six data-sets: A **Twitter tweets** data-set, a **Product Attraction** reviews data-set, a **Hotel** Reviews data-set, a **Movies** Reviews data-set, a **Product** Reviews data-set and a **Restaurants** Reviews data-set. The table below shows one example from each data-set.

Table 1: Example from each data-sets we used

Data-set	Example Instance
Twitter tweets	و الله حرام و الله موتوه لشعب الاردني من وين بدنا نجيب الكو من وين يا الله ارحمو من في الارض يرحمكم من في السماء و الله حرام
Product Attraction reviews	حمام الكبريت نصحوني بتجربة حمام الكبريت. يمكنكم الدخول مع مجموعة ولكني استأجرت غرفة كاملة لي. التكلفة 50 لاري ويمكنكم إضافة السكرب والمساج مقابل 20 لاري. ذهبت في نهاية الإجازة للاسترخاء من الرحلات اليومية التي قمت بها. الحمام نظيف والعاملات في منتهى التعاون لكن الإنجليزية ضعيفة.
Hotel Reviews	المكان الذي يمكنك فيه مراجعة الذات والتفكر هو كوكروبيت، غانا.... ثمة الكثير عند زيارة غانا. وعلى الرغم من الفقر الذي سوف تلاحظه على طريقك إلى بيغ ميلي باكيارد، وجدت أن الناس في غانا يملكون ثراء القلب حتى رغم العوز. بيغ ميلي باكيارد هو مكان يمكنني فيه مراجعة الذات والتفكر ....ً.
Movies Reviews	تلك الايام " مغامرة سينمائية متميزة إنتظرت عرض فيلم تلك " الايام كثيراً ، خاصة لإنه مأخوذ عن رواية للاديب الكبير فتحى غانم و تحمل نفس الاسم الجذاب والتى تعتبر واحدة من أهم اعمالة، وعندما عرفت أن من سيقوم باخراج الفيلم هو "أحمد غانم " ابن الروائى فتحى غانم ، شعرت بأنها شجاعة كبيرة منه خصوصاً ان هذا العمل هو أولى تجاربه السينمائية وهو عمل ليس سهلاً ، فتحويل عمل ادبى إلى فيلم سينمائى يحتاج لمخرج متمرس .....
Product Reviews	شكرا لكم على خدمتك
Restaurants Reviews	اللي يوصل ميلان ولا يمر هالمطعم اعتبرة خسران خسران مطعم مرتب اكل نظيف ولذيذ جداً و العاملين محترمين انا صراحة بدون مجاملة اعطي المطعم ١٠٠/٩٩،٩

The following table shows the percentage of positive and negative classes in each data-set and the unified data-set that contains all the data-sets:

Table 2: Data-sets distributions on classes

Data-set	Positive Samples	Negative Samples
Twitter tweets	(1000) 50%	(1000) 50%
Product Attraction reviews	(2073) 96%	(82) 4%
Hotel Reviews	(10775) 69%	(4798) 31%
Movies Reviews	(969) 63%	(556) 37%
Product Reviews	(3101) 72%	(1172) 28%

Restaurants Reviews	(8030) 73%	(2941) 27%
Unified data-set	(25948) 71%	(10549) 29%

We used a **10-fold cross validation** technique to evaluate our system. We calculated the mean of the 10-folds for the evaluations measures we used, **precision**, **recall**, **f-measure** and **accuracy**. For each data-set, we train a SVM and a Naïve Bias models. The following tables shows our results.

Table 3: 10-fold cross validation results on SVM classifier

SVM				
Data-set	**Precision**	**Recall**	**F-measure**	**Accuracy**
Twitter tweets	0.88	0.77	0.81	0.82
Product Attraction reviews	0.99	1.00	0.98	0.96
Hotel Reviews	0.96	0.98	0.88	0.83
Movies Reviews	0.87	0.95	0.82	0.73
Product Reviews	0.90	0.98	0.86	0.78
Restaurants Reviews	0.94	0.99	0.85	0.75

Table 3: 10-fold cross validation results on Naïve Bias classifier

Naïve Bias				
Data-set	**Precision**	**Recall**	**F-measure**	**Accuracy**
Twitter tweets	0.92	0.86	0.84	0.84
Product Attraction reviews	0.97	1.00	0.98	0.96
Hotel Reviews	0.95	0.99	0.83	0.72
Movies Reviews	0.79	1.00	0.77	0.63
Product Reviews	0.90	0.99	0.85	0.74
Restaurants Reviews	0.91	0.99	0.84	0.73

By loading all the data-sets into a one unified data-set, we tested the generated model using 10-fold cross validation. The following table shows the results:

Table 4: 10-fold cross validation results on unified data-set using SVM classifier

SVM				
Data-set	**Precision**	**Recall**	**F-measure**	**Accuracy**
Unified data-set	0.93	0.99	0.83	0.72

Table 5: 10-fold cross validation results on unified data-set using Naive bias classifier

Naïve Bias				
Data-set	**Precision**	**Recall**	**F-measure**	**Accuracy**
Unified data-set	0.92	0.99	0.83	0.72

Conclusion:

The results we obtained were on average around 85%. We classify positive and negative sentences. But sentiment analysis don't stop at classifying positive and negative opinions. An advanced sentiment analysis system will classify the emotional state of user utterance, such as angry, happy, etc. For Arabic, there is no data-set that is labeled by the emotional state of the speaker. Our future works include a system that can detect the emotional state of the speaker. This problem of lack resources affect all the NLP and machine learning tasks, since without data, we can't to anything. We hope that is type of data will be available in the future to support Arabic language more.

References:

[1]: Introduction to sentiment analysis https://en.wikipedia.org/wiki/Sentiment_analysis

[2]: Introduction to sentiment analysis http://sentic.net/multilingual-sentiment-analysis.pdf

[3]: Twitter role in sentiment analysis http://www.dai-labor.de/fileadmin/files/publications/narr-twittersentiment-KDML-LWA-2012.pdf

[4]: Related work on sentiment analysis https://en.wikipedia.org/wiki/Sentiment_analysis

[5]: Arabic related work on sentiment analysis https://thesai.org/Downloads/Volume6No12/Paper_11-Arabic_Sentiment_Analysis_A_Survey.pdf