Journal of Engineering and Applied Sciences 14 (14): 4780-4785, 2019
ISSN: 1816-949X

Extended-ATSD: Arabic Tweets Sentiment Dataset

Gehad S. Kaseb and Mona F. Ahmed
Department of Computer Engineering, Cairo University, Giza, Egypt

Abstract: Arabic Sentiment Analysis (SA) is one of the hottest research fields and there are still many topics open. The work in this field suffers from the lack of publicly available datasets and lexicons, however, there has been a lot of researches on SA in English. This study partially contributes by presenting a new annotated dataset, Arabic Tweets Sentiment Dataset (ATSD). The study will first detail, the process of collecting the data from Twitter for Egyptian and Saudi dialects. The gathered Tweets are classified as objective, subjective positive, subjective negative and subjective neutral. The study also discusses the process of filtering, pre-processing the dataset and annotating the Arabic text in order to build a big sentiment analysis dataset in Arabic. Determining sentiment expressed in a Tweet is not an easy task and depends on subjective judgment of human annotators. An analysis is made in order to adjust the best number of raters for new datasets annotation. The study then provides some modifications on a previous popular dataset called Arabic Sentiment Tweets Dataset (ASTD). It also combines both datasets into a collective dataset called extended ATSD. A detailed discussion of the full process adopted on the three datasets is presented. All the datasets (ATSD, Mini-ASTD and Extended-ATSD) built in this research are publicly available for academic use.

Key words: Arabic sentiment analysis, classification, datasets, machine learning, SVM, classified

INTRODUCTION

The Arabic language is a collection of different variants. Arabic dialects differ from Modern Standard Arabic (MSA) the formal written standard variant which already by itself has its own morphological and lexical complexity. This study is concerned with the most common dialect which is the Egyptian Arabic that covers the Nile valley (Egypt and Sudan) and it also handles Saudi dialect. Twitter is one of the most popular micro-blogging services, thus, a very large number of Arabic Tweets are sent every day including opinions about news, political decisions or any other trending topics. Twitter has a huge number of Arabic users who mostly post and write their Tweets using the Arabic language. Data collected from Twitter is highly unstructured and extracting useful information from Tweets is a challenging task.

Sentiment Analysis (SA) is the study of people's comments, reviews and opinions about a specific object such as an event, an item, a topic, a news feed, a mobile application or individuals. SA can be done on three levels mainly document level, sentence level and feature or aspect level. The study focuses on the sentence level SA in Tweets. The three main approaches of Sentiment Classification (SC) are lexicon-based, Machine Learning (ML) and Hybrid approaches. The actual work in the field of Arabic SA has come to light recently. The interested reader is refers to the following surveys (Abdulla *et al*., 2014; Al-Kabi *et al*., 2013; Kaseb and Ahmed, 2016; Korayem *et al*., 2012) to learn more about the work in this field. Although, much research work has been trying to make significant contributions to the field of Arabic SA, since, its rather shy start a decade ago, few of this has actually provided published datasets to be used as benchmarks for further research. Most researches applied their proposed approach on their own dataset which isn't further made public to others. This factor represents a major hindrance in the face of researchers in this field who want to prove the strength of their work by comparing their results with other work results on known benchmarks. The main target of this paper is to introduce a new dataset to be used for future research in Arabic SA also a modification over an existing well-known dataset, the ASTD is presented and a third dataset that represents the combination of the two datasets mentioned above is also made public. A four classes classifier is built using each of the three datasets and the results are reported.

Literature review: There are numerous researches on SA and a variety of approaches have been developed. English has the greatest portion of work in this field while work is somewhat limited for other languages including Arabic. This study presents some work in Arabic SA field.

Corresponding Author: Gehad S. Kaseb, Department of Computer Engineering, Cairo University, Giza, Egypt

While some of the work done in Arabic SA use the ML approaches, others used the lexicon based approach. Shoukry and Rafea (2012) worked on a dataset that contains Egyptian dialect and MSA Tweets and consists of 1,000 Tweets (500 positive and 500 negative). They used corpus-based approach where SVM and NB were used for polarity classification. The results showed that SVM outperformed NB in SA with an accuracy of 72.6%. Their work lacks handling the neutral cases and exploits a small corpus.

Abdulla *et al.* (2013) worked on a dataset that contains Jordanian dialect and MSA Tweets and consists of 2,000 Tweets (1,000 positive and 1,000 negative). They investigated the two main approaches (corpus-based and lexicon-based) of SA for Arabic corpus. A lexicon was constructed from 300 seeds and augmented in three phases to examine the impact of lexicon's size on their tool accuracy. The results showed an accuracy of 87.2% in corpus-based and 59.6% in lexicon-based approaches.

AWATIF is a multi-genre, multi-dialect corpus for Arabic subjective SA built by Abdul-Mageed and Diab (2012). It consists of about 2855 sentences of news wire stories, 5342 sentences from Wikipedia talk pages and 2532 threaded conversations from web forums. In annotating the corpus they used two different procedures, one that used untrained annotators via. crowd sourcing technologies to give a coarse sentiment label (positive, negative or neutral) to each sentence and other that used annotators trained with some linguistic background to label each sentence. The researchers also, manually created an adjective polarity lexicon that covers nearly 4,000 Arabic adjectives.

Opinion corpus for Arabic is a corpus of text from movie review sites by Rushdi-Saleh *et al.* (2011). They also translated the corpus into English. The corpus consists of 500 reviews, half negative and half positive. They performed standard pre-processing including: tokenization, removing stop words, stemming, filtering tokens whose length was <2 characters, correcting spelling mistakes and deleting special characters. The researchers used SVM and NB classifier. After several experiments using different N-grams models, the obtained results with bi-grams and tri-grams were very similar to uni-grams. In all cases the stemming process gets worse results except when using SVM on the English corpus. So, for the Arabic corpus, removing stemmer always improves the results. Using SVM, they achieved an F-score of 90% in Arabic without stemming and 88% in English with stemming.

Nabil *et al.* (2015) worked on a dataset that contains Egyptian dialect and MSA Tweets and consists of 10 K Tweets (6,691 Objective (OBJ), 1,684 negative (NEG), 799 Positive (POS) and 832 Neutral (NEU)), the dataset was called ASTD (Arabic Sentiment Tweets Dataset). It was collected using trending topics and was annotated manually. They adopted various machine learning approaches in their experiments. The best recall that they got was 69% by SVM. However, this study did not mention any pre-processing or cleaning steps also the dataset contains small number of subjective tweets.

Large-scale arabic book review corpus (Aly and Atiya, 2013) consists of 63,257 book reviews which are annotated with a scale from 1-5 where rates 1 and 2 can be considered as negative, rate 3 as neutral and rates 4 and 5 as positive. The researchers attempted to make the dataset as huge as possible. The 5 sentiment polarity classes contain 2,939, 5,285, 12,201, 19,054 and 23,778 reviews, respectively.

MATERIALS AND METHODS

Dataset collection and annotation: This study outlines the description of the new dataset, Arabic Tweets Sentiment Dataset (ATSD) which is collected and manually annotated. Building ATSD involves two main phases.

Dataset collection: The first step is collecting over 5,000 Arabic Tweets. The focus is on the top trending hash-tags in Egypt in the year 2017 exactly in February. Table 1 shows the used hash-tags in collecting the dataset with their translations.

The second step is performing some pre-processing to clean up unwanted content like HTML. Also, the gathered information sometimes contains Tweets in different languages making the information exceptionally uproarious. So, the third step is filtering out non-Arabic Tweets. The fourth step is removing duplicate tweets

Table 1: The hash-tags used to build ATSD

Arabic hashtag	Translation
#عيد_الحب	Valentine's day
#مصر	Egypt
#صراحه	Frankness
#البدله_الحمرا	The red suit
#قررت_اني_ابطل	I decided to stop
#اخبار_دلوقتي	News now
#مطلوب_من_وزير_التعليم	Which is required from the minister of education

Tweet	Label
الجمله رقم <375> _: (عسى الله أن يأتيني بهم) عندما تستبشر بالله خيرا فإن الله لايعطيك بقدر أملك به بل يزيدك من كرمه فيعقوب عاد له من غاب من بني...	Positive
الجمله رقم <685> سعر #الذهب اليوم في مصر الاثنين مقابل #الجنيه المصري	Objective
الجمله رقم <2500> _: #عيد_الحب لدينا اجود انواع #ثمور #الاحساء #خلاص جامبو حبة كبيرة التوصيل لجميع مناطق المملكةوالخليج ع الرقم التالي ...	Neutral
الجمله رقم <1193> الحكم الاستبدادى #ايه_اللي_جابنا_ورا	Negative
الجمله رقم <2293> اجعنك ياخوان الحبمرض لاينز ومعاه القلب خواني مانبيله شيعمليه وايقص الضي،، اجعنك ياخوان مديدفي ثلاجه ليله عيد	Unknown

Fig. 1: ATSD sample Tweets

(re-Tweet) and the Tweets that contain only links or even most of them are links. It ended up with 2,500 Arabic Tweets.

Dataset annotation: A human read the 2,500 Tweets and removed 473 Tweets that were dirty data (e.g., Tweets that contain a lot of hashtags or not completed Tweets) that can't be annotated. Then 2027 Tweets are assigned to six different raters for annotation. The raters are native Arabic speakers who represent samples from different combinations of age, gender, city and education. Because the point of view upon which a sentence is built influences, whether it is taken as positive or negative, annotators were asked to label sentences based on their background regardless of whether the Tweet is ambiguous. So, there are different opinions for the same Tweet; only 450 Tweets have consistent opinion from all raters.

Inter Rater Reliability (IRR, also called inter-annotator agreement) is the degree of agreement among raters. It gives a score of how much homogeneity or consensus, there is in the ratings given by judges and it is one of the aspects of test validity. Table 2 shows the IRR for ATSD. Finally, the following steps are adopted on the annotated Tweets till reaching the final dataset:

First: Removing Tweets which get more than two unknown opinion.

Second: Removing about 200 Tweets which are similar but differ only in writing style.

Third: Retaining Tweets that have at least three raters who provided the same opinion given that the other raters do not agree on the contrary but removing others that have three raters with the same opinion and the other three raters with a contrasting opinion.

Table 2: IRR between different raters for ATSD

Inter rater reliability/Agreement (raters)	Values (%)
IRR-6	18.50
IRR-5	20.1-25.8
IRR-4	24.57-32.07
IRR-3	29.96-39.37
IRR-2	44.65-61.67

Table 3: Datasets statistics

Variables	ASTD	Mini-ASTD	ATSD	Extended-ATSD
Objective Tweets	6,691~67%	3,238~59%	292~18.5%	3,530~50%
Subjective negative Tweets	1,684~17%	1,328~24%	573~36.5%	1,901~27%
Subjective positive Tweets	799~8%	564~10%	373~23.5%	937~13%
Subjective neutral Tweets	832~8%	390~7%	324~20.5%	714~10%
Total number of Tweets	10,006	5,520	1,562	7,082

It ended up with 1,562 Arabic Tweets dataset. Figure 1 shows examples of Tweets with different labels in ATSD. Table 3 shows ATSD statistics.

Dataset raters analysis: There are six raters who are initially hired to annotate ATSD. The selection of six specifically is made initially at random with the aim of getting enough confirmation on the decision of the class to which each Tweet belongs. They started with the ATSD with an initial size of 2027 Tweets. Table 4 shows the specification of the different annotators that are all Egyptian native Arabic speakers.

Table 4: Specication of the raters

Rater ID	Genders	Age	Education	City
R1	Female	26	Engineering	Giza
R2	Male	22	Commerce	BeniSuef
R3	Female	24	Pharmaceuticals	Beheira
R4	Female	22	Commerce	Giza
R5	Female	42	Engineering	Monufia
R6	Female	22	Geography	Cairo

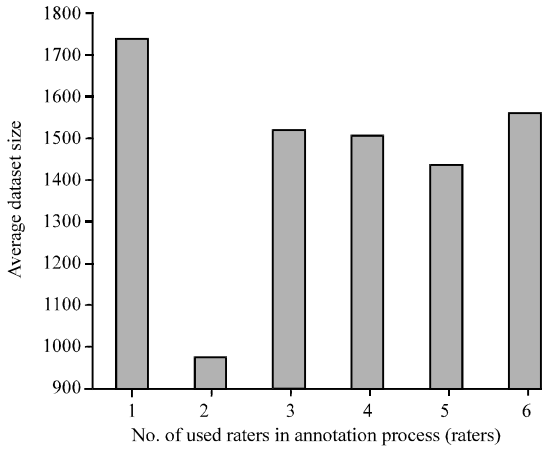

Fig. 2: Dataset size vs. raters number

An analysis is made with the aim of arriving at the most convenient number of raters to help researchers in their future dataset building work. The analysis depends on annotating the dataset with different numbers of raters to try to arrive at the best choice of the actual required number of raters. The result shows that the resultant dataset size depends on the selected number of raters. Six iterations are done for this purpose the iteration number refers to the number of raters included in this iteration, e.g., iteration one corresponds to having only one rater.

In each iteration, all possible combinations of the six raters are made and the resultant average datasets sizes are calculated. Each combination of raters gives different dataset size, however, the results are close to each other. Figure 2 shows the six iterations with the corresponding average dataset size. Obviously, using one rater gets maximum dataset size but this case is not significant as it expresses only one viewpoint in building the new dataset. Using two raters gets the minimum dataset size as having conflicting opinions will discard items and minimizes the dataset size. Using three, four or five raters gives close dataset size. So, for the researchers who want to build new dataset it is recommended according to these results to use three different raters. Using more than three raters will be useless and consume time and effort especially for big dataset size.

Mini-ASTD: The second dataset is ASTD (Nabil *et al*., 2015) which consists of 10,006 Tweets written in MSA as well as Egyptian colloquial Arabic. It incorporates four tags: objective, subjective positive, subjective negative and subjective mixed. Table 3 shows ASTD statistics.

The main issue in ASTD is that most of the Tweets are non-subjective, this issue limits the dataset usage. Suspecting incorrect labeling of the ASTD dataset, work was done on cleaning the ASTD through the following steps (steps one to three are done automatically whereas those from four to eight are done manually):

First: Removing full duplication.

Second: Removing tweets which have no meaning that are just stop words or famous names (e.g., مصطفي_محمود and القاضي كريم اسماعيل. #عثمان_بن_عفان). They were detected with the aid of Named Entities Recognizing (NER) Gazetteer lists. Using ANERGazet (14) which is a collection of three Gazetteers (Location Gazetteer, Person Gazetteer and Organizations Gazetteer) (Benajiba *et al*., 2007).

Third: Removing partially duplicated Tweets. This step is achieved simply by removing duplication after the preprocessing phase. An example, of partial duplication is shown below:

عرض الاسبوع 40 الف متابع #شوار عنا #فن_تتقنه_النساء
#ذلك_الشخص #شي_ودك_تجربه #بوح #درر #عجبني #حلو #مكه
2063421]] OBJ

عرض الاسبوع 40 الف متابع #شوار عنا #فن_تتقنه_النساء
#ذلك_الشخص #شي_ودك_تجربه #بوح #درر #عجبني #حلو #مكه
OBJ
...

Fourth: Removing meaningless Tweets, e.g.

وقت #عبدالحليم_حافظ 3 OBJ

إحديث الساعة OBJ

Fifth: Removing Tweets that contain a lot of hashtags with no real message (they are considered spam).

النصر #متصدرلاتكلمني #بعترفاني #عيدالحب # :9"
#غردبصورة #اليمن #حضرموت #تصويري #دوري_جميل
#الهلال " OBJ

الكويت #تبي #السعوديه #الرياض #سوريا #الهلال #
#السعودية #النصر #غرد_بصوره OBJ

Sixth: Removing Tweets which are not completed (the meaning is not clear due to missing words), e.g.

يمكنكم متابعة البث الحي عبر OBJ

أكيد هم اخوان او ... OBJ

Seventh: Removing Tweets which contain one or two words only but with no meaning, e.g.

المتجولون OBJ

باقي_على #رمضان# OBJ

Eighth: Removing Tweets which have obviously wrong sentiment as shown in Table 5.

After the previous steps, the dataset ended up with 8,800 Tweets from 10,000 (Ignore 12% of meaningless/inexpedient Tweets). The remaining Tweets are annotated again. IRR between the old annotation and the new one is 61.7%. Finally, only the Tweets which have consistent annotations are accepted. Mini-ASTD ended up with 5,520 Arabic Tweets. Table 3 shows Mini-ASTD statistics. Mini-ASTD solves the biasing towards the objective class to a great extent. It is about 67% in ASTD and is reduced to about 59% in Mini-ASTD. However Mini-ASTD is about half the size of ASTD, yet Mini-ASTD is more robust for further uses.

Extended-ATSD: This research presents two datasets, ATSD and Mini-ASTD, each one has an advantage over the other. Mini-ASTD is about 3.5 times ATSD in size. ATSD is more balanced than the Mini-ASTD concerning the class labels. So, to take the advantage of both another dataset called Extended-ATSD is presented. Extended-ATSD combines both ATSD and Mini-ASTD. Table 3 shows Extended-ATSD statistics. It reaches 7K Arabic Tweets where about half of them are objective and the rest are subjective. Extended-ATSD can be used as a four polarities Arabic SA dataset (~7K) or as subjective Arabic SA dataset (~3.5K) by removing the objective class. It can be used as positive-negative dataset (~3K) and also as subjective-objective dataset (~7K) (~50%-50%). ATSD, Mini-ASTD and Extended-ATSD are publicly available for academic use.

Table 5: A sample of wrong classified tweets in ASTD

Actual	Expected	Tweets
OBJ	NEG	ماحدش عنده #ضحكة سلف #يشعر بالحزن
POS	NEG	و لعل أسوء ما في هذا اليوم غيابك يا #ياسم
OBJ	POS	مبدع دائما يا عم الشيخ
OBJ	POS	#اعترافات_ليلية احب المحلوه

Classification: This study outlines the methodology used in this research. In the adopted experiments the data is used as it is without any pre-processing, to feed the classifiers either in the training or the testing phase. From the study, comparison and analysis of the different proposed methodologies for SA it was observed that SVM yields the best performance over other ML classifiers (Kaseb and Ahmed, 2016).

SVM gained this prominent position because of its advantages which can be summarized as follows; It is robust in high dimensional spaces. All analyzed features are considered relevant. SVM is robust when there is a sparse set of samples. Finally, most text categorization problems are linearly separable, so, they are good candidates for SVM (Saleh *et al*., 2011). Accordingly, SVM was used with TF-IDF feature vector to provide classification into 4 classes: OBJ, POS, NEG and NEU.

RESULTS AND DISCUSSION

In order to evaluate the performance, four metrics are used: accuracy, precision, recall, F-measure. To explain these measures it is generally assumed for a binary classification problem that there is a "positive" class and a "negative" one. Precision calculates the ratio of the true positives to the total number of positives predicted by the classifier. The higher the precision is the more accurate the prediction of the positive class. On the other hand, recall divides the true positives by the total actual positives that belong to that class. A high recall means a high number of Tweets from the same class are labeled with their true class. F-measure (F-score or F1-score) is the weighted harmonic average of precision and recall. As for the accuracy, it simply reports the ratio of the correctly classified Tweets to the total number of Tweets regardless of their class.

The algorithm is run 10 times using a random seed to shuffle and partition the dataset into training set, 80% and test set, 20%. This randomness changes the training and test sets in every iteration which means that evaluation is done using 10 combinations of the training and test sets. Tables 6 and 7 show accuracy, precision, recall and F1-score achieved over the different datasets.

Table 6: ASTD vs. Mini-ASTD accuracy, precision, recall and F1-score

Variables	Accuracy	Precision	Recall	F1-score
ASTD	64	58.3	63.9	60.2
Mini-ASTD	64.7	61.3	64.6	62.4

Table 7: ATSD vs. extended-ATSD accuracy, precision, recall and F1-score

Variables	Accuracy	Precision	Recall	F1-score
ATSD	53.8	54.2	53.9	53.7
Extended-ATSD	62.4	60.3	62.4	61.1

CONCLUSION

Three datasets (ATSD, Mini-ASTD and Extended-ATSD) are presented to build a big SA dataset for Arabic language. ATSD is gathered from Twitter then manually annotated with four polarities (OBJ, POS, NEG and NEU). This work recommends using three different raters to annotate any new dataset. Using more than three raters will be useless and consume time and effort especially annotations and using two leads to the smallest resulting dataset size because of the conflicting opinions.

Mini-ASTD is a small filtered and cleaned dataset from ASTD that is manually annotated. Extended-ATSD is the combination of ATSD and Mini-ASTD. This work uses SVM classifier and TF-IDF feature extraction. The resultant accuracy reached 64.7 for Mini-ASTD, 64 for ASTD, 62.4 for extended-ATSD and 53.8 for ATSD. The results show that Mini ASTD excels over the original ASTD in the different performance measures because of the enhancement done on the dataset. ATSD gets the lowest measurements as it is the smallest in size, so, the classifier is not trained well as is done with a larger dataset, so, both Mini ASTD and ATSD are combined to get the final proposed dataset, extended ATSD, trying to get the best of the two in the large size and balanced dataset. There are slight differences in the evaluation measures between Mini-ASTD and extended-ATSD because the work focuses on presenting the datasets without exploiting any special pre-processing that can affect the results which will be considered in the next steps. The intended future research is to build a hybrid methodology for Arabic SA using Extended-ATSD and to create a large scale lexicon from this corpus.

REFERENCES

Abdul-Mageed, M. and M.T. Diab, 2012. AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'14), May 26-31, 2014, European Language Resources Association, Reykjavik, Iceland, pp: 3907-3914.

Abdulla, N., M. Al-Ayyoub and M.N. Al-Kabi, 2014. An extended analytical study of Arabic sentiments. Intl. J. Big Data Intell., 1: 103-113.

Abdulla, N.A., N.A. Ahmed, M.A. Shehab and M. Al-Ayyoub, 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. Proceedings of the 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT'13), December 3-5, 2013, IEEE, Amman, Jordan, ISBN:978-1-4799-3676-2, pp: 1-6.

Al-Kabi, M., N.M. Al-Qudah, I. Alsmadi, M. Dabour and H. Wahsheh, 2013. Arabic/English sentiment analysis: An empirical study. Proceedings of the 4th International Conference on Information and Communication Systems (ICICS'13), April 23-25, 2013, ACM, Irbid, Jordan, ISBN:978-1-4503-1327-8, pp: 23-25.

Aly, M. and A. Atiya, 2013. Labr: A large scale Arabic book reviews dataset. Proceedings of the 51st Annual Meeting on Association for Computational Linguistics, August 4-9, 2013, Association for Computational Linguistics, Sofia, Bulgaria, pp: 494-498.

Benajiba, Y., P. Rosso and J.M. Benediruiz, 2007. Anersys: An arabic named entity recognition system based on maximum entropy. Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing, February 18-24, 2007, Springer, Berlin, Heidelberg, Germany, ISBN:978-3-540-70938-1, pp: 143-153.

Kaseb, G.S. and M.F. Ahmed, 2016. Arabic sentiment analysis approaches: An analytical survey. Int. J. Scient. Eng. Res., 7: 712-723.

Korayem, M., D. Crandall and M. Abdul-Mageed, 2012. Subjectivity and Sentiment Analysis of Arabic: A Survey. In: Advanced Machine Learning Technologies and Applications, Hassanien, A.E., M.S. Abdel-Badeeh, R. Ramadan, and K.T. Hoon (Eds.). Springer, Berlin, Germany, ISBN:978-3-642-35325-3, pp: 128-139.

Nabil, M., M. Aly and A. Atiya, 2015. ASTD: Arabic sentiment Tweets dataset. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, September 17-21, 2015, Association for Computational Linguistics, Lisbon, Portugal, pp: 2515-2519.

Rushdi-Saleh, M., M.T. Martin-Valdivia, L.A. Urena-Lopez and J.M. Perea-Ortega, 2011. Bilingual experiments with an Arabic-English corpus for opinion mining. Proceedings of the International Conference on Recent Advances in Natural Language Processing, September 12-14, 2011, Hissar, Bulgaria, pp: 740-745.

Saleh, M.R., M.T.M. Valdivia, L.A.U. Lopez and J.M.P. Ortega, 2011. OCA: Opinion corpus for Arabic. J. Am. Soc. Inf. Sci. Technol., 62: 2045-2054.

Shoukry, A. and A. Rafea, 2012. Sentence-level Arabic sentiment analysis. Proceedings of the 2012 International Conference on Collaboration Technologies and Systems (CTS'12), May 21-25, 2012, IEEE, Denver, Colorado, ISBN: 978-1-4673-1381-0, pp: 546-550.