Framework for Sentiment Analysis of Arabic Text

Latifah Almuqren
Department of Computer Science,
Warwick University,
Coventry, UK.
L.Almuqren@warwick.ac.uk

Alexandra I. Cristea
Department of Computer Science,
Warwick University,
Coventry, UK
A.I.Cristea@warwick.ac.uk

ABSTRACT

This paper analyses *challenges*, and provides a *model* and a *framework* for mining Arabic tweets to measure *customer satisfaction* toward telecom companies in Saudi Arabia, to predict the ratio of *customer churn* and overcome the specific challenges with the semantic sentiment analysis of Arabic text.

Keywords

Semantic Sentiment Analysis (SSA), Arabic, Twitter, Sentiment

1. INTRODUCTION

Enhancing customer satisfaction is a popular topic in marketing literature, as much research correlates customer satisfaction with customer loyalty [6]. Customer satisfaction is attained by examining customer expectations toward the products of a company [8]. *Customer churn* is defined in the telecom field as transferring from one Telecom Company to another [14]. Customers who are satisfied with company services make a company more profitable, because the cost of attracting new clients is five times more than retaining customers [6]. Traditionally, *customer satisfaction* has been measured through customer interviews and questionnaires, but these cannot measure customer satisfaction in real time [12].

Twitter is a widely popular messaging service categorised as a micro-blogging website [16]. Messages in Twitter are embedded in so-called *tweets*, which are individual, unstructured text messages with a limit of 140 characters [16]. By mining these tweets, a database of emotions can be created, to analyse the sentiments and related subjective contexts of micro-blogging conversations [22]. Analysing subjective contexts and emotions can help in creating real time analytics or mining information on public opinion and emotions about products, leadership, decisions, cultures and events [12].

The Arabic language is quite challenging for SSA interpretation [13], due to the variety of forms in Arabic language, such as Modern Standard Arabic (MSA) and the informal language, or colloquial Arabic [7], and the structure of the language, such as the direction of writing from right to left.

However, *customer satisfaction* based on real time methods for Saudi Arabian companies is an understudied area, for reasons that are further described in this paper. Thus, this study examines microblogging site mining techniques for the purpose of capturing user satisfaction toward telecom companies in Saudi Arabia, and how can we use that data to give recommendations to these companies.

In the current paper, we focus on investigating three questions:

RQ1. What are the variables influencing customer satisfaction and affect customer churn?

RQ2. What social media data mining techniques are appropriate for capturing customer satisfaction toward telecom companies in Saudi Arabia from microblogging sites?

RQ3. What are the challenges for SSA tools for Arabic tweets?

The remainder of the paper is structured as follows. Section 2 presents our *customer satisfaction model* and related work. Section 3 explains our suggested *research framework*. Section 4 summarizes the *challenges in Arabic Semantic Sentiment Analysis* (SSA). Finally, Section 5 provides conclusions and future work.

2. NEW MODEL AND RELATED WORK

From a literature analysis of 50 papers on predictions of customer satisfaction based on Twitter, and of Twitter analysis for Arabic text, we have collected all variables that linked Twitter and user satisfaction, which were shown, in these different studies, to influence customer satisfaction and affect *customer churn*, from which we created our **customer satisfaction model** (Figure 1).

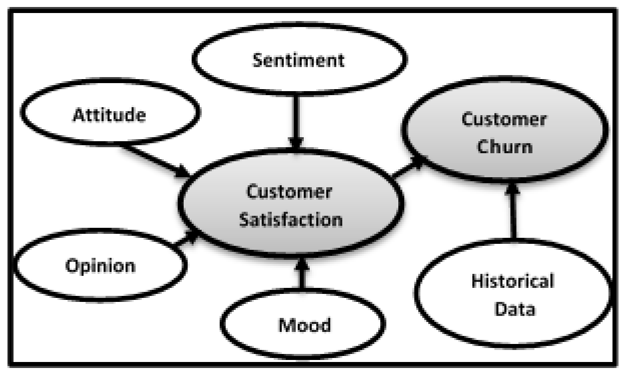

Figure 1: Our customer satisfaction model

Mood

Bollen et al. [10] examined the influence of public mood on the closing value of the Dow Jones Industrial Average (DJIA). They used two mood tracking tools, Opinion Finder and Google-Profile of Mood States (GPOMS), to analyse the text of daily tweets on Twitter. Then, they applied Granger Causality analysis to predict the correlation between a certain mood and DJIA value. The results indicated that not all changes in the public mood matched DJIA value shifts, but a 'calm' mood could predict DJIA values.

Opinion, Attitude and Sentiment

Salampasis et al. [23] analysed consumers' behaviour about food products, using their micro blogging messages (i.e., tweets) to monitor and analyse consumer opinion, attitude and sentiments expressed in shared posts and comments. Results showed that the success of branding required monitoring sentiments for a long period of time, because these sentiments do not change quickly.

Semantic Sentiment Analysis

Our study plans to use Semantic Sentiment Analysis (SSA) of Arabic tweets to measure customer satisfaction. Collines et al. [11] measured public transport rider satisfaction toward transit system services using the riders' tweets on Twitter. This research helped the transit system to improve the service quality and safety monitoring by adding more personnel. The survey analysed the tweets of riders along the Chicago Transit Authority (CTA) rapid transit system using a Sentiment Strength Detection Algorithm (SentiStrength), to detect rider sentiments in real time.

Mostafa [18] analysed 3,516 tweets to measure consumer sentiment toward brands such as Nokia, T-Mobile, IBM, and DHL, using a predefined lexicon including around 6,800 seed adjectives with known orientation. Results indicated a generally positive consumer sentiment toward several famous brands.

Historical Data

Many studies have focused on using historical data, using data mining to create a prediction model [14]. Also, Kampakis & Adamides [15] predicted the score of football matches using Twitter mining and historical data, with a high accuracy of 75%.

3. USER SATISFACTION FRAMEWORK

Based on previous studies of Twitter mining, as well as on the lack of tools and annotated corpuses for Arabic, we suggest a new study with the following phases (Figure 2):

1. **Data Collection:** Build a corpus of Arabic SSA messages from the Twitter semantic (search) API [1,12,21], with Arabic native speakers searching and adding annotations.
2. **Building an In-Domain Arabic Lexicon**: Use a corpus-based approach to build the lexicon, using all data from the same domain. The classified tokens (breaking Arabic expressions down into smaller tokens) and lemmas (each word reduced to a citation form) may be used as seeds for an automated machine learning algorithm, after weighing term-frequency (to judge and include the additional words from the *Saudi dialectal Arabic* (SDA)), and then finalise the lists of positive, negative and neutral words [1,16]. A lexicon seed comprises a list of words from the SDA having identical semantic, lexical and semantic-lexical relationships for positive, negative and neutral emotions [2, 3, 16]. A seed is constructed from frequently used words in Twitter and never used words in Twitter (via expansion algorithms) [2, 16].

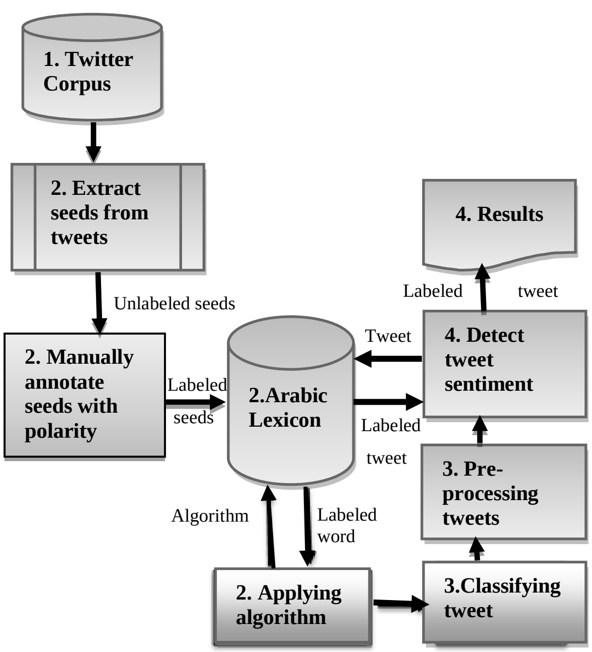

Figure 2: The User Satisfaction Capturing Framework

3. **Pre-Processing Tweets:** Normalize and tokenize all tweets. Tokens are stemmed via a revised n-gram approach [7, 16].
4. **SSA**: Detect customer satisfaction using SSA for each tweet, applying classifiers with proven high accuracy for Arabic text: Support Vector Machines, KNN, Naïve Byes [3, 4, 12].

4. CHALLENGES AND SOLUTIONS

Most SSA tools are for English. These need a lot of adjustment to function with Arabic. To develop SSA tools for Arabic text requires an understanding of the unique Arabic internal structure, nature, terms and linguistic rules, plus dialectical and colloquial differences in different Arabic regions. Each form has its own syntax and vocabulary, which makes building an Arabic lexicon difficult [5]. Additionally, different words may have the same meaning in different dialects. For example, "a lot" in the Moroccan dialect is (باهي), but it is (صافي) in Libyan [19].

Arabic script properties also pose challenges, such as:

- **Diacritization:** A word may have different meanings, based on the small diacritical marks above or under letters. For example (شعر) could mean "hair" with a small diacritical mark above its first letter, or it could mean "poem" with a small diacritical mark below that letter [20].
- **Negation:** In English, negative words often use a prefix, such as un- or im-, etc. (e.g., impossible vs. possible). In Arabic, however, a negative word is preceded by a separate term. For example, *unhelpful* (غير مفيد) vs. *helpful* (مفيد). This causes a problem, as the sentence-level classifier treats the negation as two separate words. Azmi & Alzanin solved this problem by joining the words in a preprocessing phase [9].
- **Spelling Errors:** Mixing up small dialectical marks and long vowels leads to errors: e.g., *you* (انتي) should be spelled (انت).

Azmi & Alzanin and Saif et al. [9, 22] proposed a revised n-gram to correct misspelled words and improve classification. If they couldn't find a word in a vocabulary table, they searched for similarities in that table. If two words were at least 70% similar, they counted them as referring to the same notion, such as the terms *succccess* (نجاااح) and *success* (نجاح) (Figure 3).

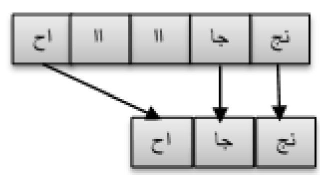

Figure 3: A revised n-gram approach

5. CONCLUSIONS

Due to increasing global competition, companies want to know *customer sentiments*, to improve customer care. This research provides a *customer satisfaction model* to compute *customer churn* (answering RQ1), a *framework to capture user satisfaction* toward telecom companies (answering RQ2) and examines some *challenges and their potential solutions* (answering RQ3) in using SSA for Arabic text, due to the unique nature of Arabic. To meet the challenges related to Arabic script properties, this study proposes solutions in the pre-processing stage, such as normalization, revised n-gram or contextual rules. These solutions are expected to greatly enhance classifier accuracy. Additional research in this area should be conducted using these suggested approaches.

6. REFERENCES

[1] Abdulla, N., Majdalawi, R., Mohammed, S., Al-Ayyoub, M. and Al-Kabi, M. Automatic lexicon construction for arabic sentiment analysis. *in Future Internet of Things and Cloud (FiCloud)*, (Barcelona,Spain, 2014), IEEE , 547-552

[2] Abdul-Mageed, M., Diab, M. T. and Korayem, M., Subjectivity and sentiment analysis of modern standard Arabic. *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Shortpapers*,(Stroudsburg, PA, USA, 2011), Association for Computational Linguistics, 587–591.

[3] Abdul-Mageed,M., Kübler, S. and Diab, M. SAMAR: A system for subjectivity and sentiment analysis of Arabic social media. *Association for Computational Linguistics*, 28(1).20-37.

[4] Ahmed, S., Pasquier, M. and Qadah, G., Key issues in conducting sentiment analysis on Arabic social media text. *in 9th International Conference on Innovations in Information Technolog (IIT)*, (Abu Dhabi,UAE,2013) IEEE, 72-77 .

[5] Albraheem, L. and Al-Khalifa, H. S., Exploring the problems of sentiment analysis in informal Arabic. *In Proceedings of iiWAS*,(Bali,Indonesia, 2012),ACM, December, 415-418

[6] Allola, A. 2011. Measuring the quality of services provided by Jawwal Co. from the viewpoint of customers in the provinces of Gaza Strip. Unpublished Master thesis, University of Alazhar, Gaza, Palestine.

[7] Al-Twairesh,N., Al-Khalifa, H. and Al-Salman, A., Subjectivity and sentiment analysis of Arabic: Trends and challenges.*in 11th International Conference on Computer Systems and Applications*,(Doha, Qatar, 2014), IEEE/ACS, 148-155.

[8] Athanassopoulos, A. D. and Iliakopoulos, A. Modeling customer satisfaction in telecommunications: Assessing the effects of multiple transaction points on the perceived overall performance of the provider.*Production and Operations Management*, 12(2), 224-245, 2003

[9] Azmi, A. M. and Alzanin, S. M. Aara'–a system for mining the polarity of Saudi public opinion through e-newspaper comments. *Journal of Information Science*, 40(3), 398-410, 2014

[10] Bollen, J., Mao, H. and Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8, 2014

[11] Collins, C., Hasan, S. and Ukkusuri, S. V. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. *Journal of Public Transportation*, 16(2), 2, 2013.

[12] Duwairi, R.M. and Qarbaz,I., Arabic Sentiment Analysis Using Supervised Classification. *in International Conference on Future Internet of Things and Cloud (FiCloud)*, (Barcelona ,Spain,2014), IEEE Computer Society, 579-583.

[13] Farghaly A. and Shaalan, k. Arabic natural language processing: Challenges and solutions. *Transactions on Asian Language Information Processing*,8(4), 1-22, 2009.

[14] Hung, S. Y., Yen, D. C. and Wang, H. Y. Applying data mining to telecom churn management. *Expert* Systems with Applications, 31(3), 515-524, 2016

[15] Kampakis, S. and Adamides, A. Using Twitter to predict football outcomes. *ArXiv preprint arXiv:1411.1243*, 2014, Retrived January 20, 2016 from Cornell University Library : http://arxiv.org/abs/1411.1243

[16] Mahyoub, F. H. H. , Siddiqui, M. A. and Dahab, M. Y. Building an Arabic sentiment lexicon using semi-supervised learning. *Journal of King Saud University Computer and Information Sciences*, 26(4), 417–424, 2014.

[17] Marcus, A., Bernstein, M. S., Badar,O., Karger, D. R., Madden, S. and Miller, R. C. Processing and visualizing the data in tweets. *SIGMOD Record*, 40(4), 21-27, 2011.

[18] Mostafa, M. M. More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10), 4241-4251, 2013.

[19] Mourad,A. and Darwish,K., Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. in *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*,(Atlanta, Georgia, United States, 2013), Association for Computational Linguistics, 55–64.

[20] Mustafa, M., AbdAlla, H. and Suleman, H. Current approaches in Arabic IR: A survey. In G. Buchanan, M. Masoodian and S. Cunningham, ed., *Digital Libraries: Universal and Ubiquitous Access to Information.* Springer Berlin Heidelberg, Bali, Indonesia, 2008, 406-407.

[21] Refaee,E., and Riesner,V., An Arabic Twitter corpus for subjectivity and sentiment analysis. in *Proceedings of Language Resources and Evaluation Conference 2014*, (Reykjavik, Iceland, 2014), 2268 -2273.

[22] Saif, H. He, Y. and Alani, H., Semantic sentiment analysis of Twitter. in *Proceedings of the 11th International Conference on the Semantic Web*,(Boston, Massachussetts, United States, 2012), Springer-Verlag Berlin, 508-524.

[23] Salampasis, M., Paltoglou, G. and Giachanou, A. Using social media for continuous monitoring and mining of consumer behaviour. *International Journal of Electronic Business*, 11(1), 85-96, 2013