Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Kurdish Language Sentiment Analysis: Problems and Challenges

Miran Hama Saeed Mohammed Amin[1], Omar Al-Rassam[2], Zhenar Shaho Faeq[3]

[1]Department of Software Engineering, Faculty of Engineering, Koya University, Koya KOY45, Kurdistan Region - F.R. Iraq

[2]Department of Mathematics, Faculty of Science and Health, Koya University, Koya KOY45, Kurdistan Region - F.R. Iraq

[3]Department of Software Engineering, Faculty of Engineering, Koya University, Koya KOY45, Kurdistan Region - F.R. Iraq

Article Info

Page Number: 3282 - 3293

Publication Issue:

Vol 71 No. 4 (2022)

Article History

Article Received*: 25 March 2022*

Revised*: 30 April 2022*

Accepted*: 15 June 2022*

Publication*: 19 August 2022*

Abstract

The increasing usage of blogs, social networks, and forums for sharing opinions toward a certain topic has created a vast amount of data over the internet. Therefore, Sentiment Analysis has gained great popularity among researchers and industry for analyzing the polarity of users' opinions. In recent years, Sentiment Analysis has been applied to various languages using machine learning-approach, corpus-based approach, and deep learning techniques since it is beneficial for creating an effective recommender system. Kurdish language is an Indo-European language, one of the official languages in Iraq and it is also widely used in Turkey, Iran, and Syria. Although the importance of this language which is spoken by over 40 million people, to the best of our knowledge, no research has been done regarding the challenges and problems of Kurdish sentiment analysis. Our research work is to highlight on the latest studies and examines the most important challenges of applying sentiment analysis approaches to the Kurdish language. The study includes determining each challenge in each step of sentiment analysis processing in Kurdish language. In addition, our proposed methodology that could help to address most of these challenges is to implement a hybrid approach by combining machine learning approach and lexicon-based approach to improve the proficiency of sentiment classification for Kurdish language.

Keywords- Sentiment Analysis, Machine Learning, Social Networks, Kurdish language, Kurdiz, Corpus

1. Introduction

Nowadays, one of the trending research topics in almost every field including business, marketing and health is sentiment analysis. Sentiment Analysis is the field that explores the computational study of human thoughts or opinions, emotions and attitudes toward a certain entity. In the past decade, opinion extraction has been employed in many languages, especially English language. Opinion mining can be easily achieved in English language due to the availability of resources including lexicons, dictionaries and corpora while the lack of these resources in the other languages have made challenges for researchers in order to apply accurate opinion mining [1].

Researchers have analyzed Sentiment on non-English Languages Including Spanish, Italian, German, Dutch, Chinese, Japanese, Taiwanese, Persian and Arabic. However, they have encountered common challenges due to the lack of resources such as lexica, corpora and dictionaries for those languages [2].

Sentiment analysis can be classified into three levels, document-level, sentence-level, and aspect-level. One of the most important aspects of conducting an accurate Sentiment Analysis is to have a large data set. Data sets are mainly formed from Social Media blogs and product reviews since they are crucial for business owners in order to make decisions. Moreover, sentiment analysis is not only applied on product reviews, but also to stock markets, news articles, or political debates [2].

Researchers working on sentiment analysis either generate their own data or use available datasets. Creating a new dataset has two advantages; the generated data fits the problem of the targeted analysis and ensures that no privacy laws are violated. However, the main drawbacks of using a newly created dataset are labeling which is a challenging task and it is difficult to generate a large volume of data [3]. The most common lexicon source for non-English languages is WorldNet. However, more resources need to be created to be used in sentiment analysis for many languages [2].

Recently, few researchers have conducted studies for applying sentiment analysis to Kurdish context due to the challenges of the language. These challenges need to be determined and addressed in order to get precise results in KSA.

Kurdish language users express their feelings in different ways, they do not follow standard writing rules, neither grammar syntax nor alphabet. Some users prefer to use Arabic script letters while others prefer to use Latin letters to express their feelings. Moreover, using

different forms writing in Kurdish text increase the challenge and sometimes people use English letters instead of the standard Kurdish alphabet to express their feelings [4].

In addition, lack of language resources and the diversity of the language standardization and segmentation issues are among the main challenges in Kurdish text processing [5]. In addition, dialect Diversity is considered among the most challenging issues since there are various dialects in Kurdish language [6].

Kurmanji and Sorani are the most used dialects among the others which they account for more than 75% of native Kurdish speakers [7]. Therefore, it is difficult to create a large scale corpus for Kurdish language considering all dialects.

In addition, the challenge gets bigger since the writing system of the two most used dialects are where Kurmanji dialect is written in Latin-based letters and Sorani is written in Arabic-based letters [8].

1.1 The aims of the work

The main objective of this research is to offer a comprehensive study of the most important challenges faced by researchers in the field of Kurdish sentiment analysis (KSA). The study includes determining each challenge in each step of sentiment analysis processing in Kurdish language. Moreover, the most important point in this work is to focus on all the Kurdish dialects and scripts which are Soran, Kurmanji, and the written Kurdish language in both forms alphabet and Latin. In addition, a proposed methodology is presented to address most of these challenges which could help new researchers in this field.

2. Related work

Due to the increasing usage of social media platforms and a large volume of product reviews on various e-commerce websites, sentiment analysis has become one of the most dynamic research fields in natural language processing (NLP). Over the past decade, many studies have been performed in order to address the challenges in Sentiment Analysis for many languages and increase accuracy. However, to the best of our knowledge, no studies have been done to define and address Kurdish language challenges.

Kurdish Language as a less-resourced language has similar forms, alphabets and orientation (right to left) with some other low resourced languages such as Arabic and Urdu. Therefore, most of the challenges of applying Sentiment Analysis in those languages are common.

Lack of Corpora and Sentiment lexica, Dialect Diversity and Code Switching are some major challenges of Sentiment Analysis in Arabic language [9]. [10,11,12,13] determined some

other challenges such as spelling, vocabulary, phonetics, and morphology in Arabic Language increases syntactic, semantic, and figurative ambiguity. Moreover, Real-time sentiment analysis, spam detection, morphological faults, inadequate spelling, unstructured data, and implicit meanings are also among the challenges of Arabic Sentiment Analysis. Finally, the community working on Arabic Sentiment Analysis is small.

In addition, [14] conducted a study on Sentiment Analysis of Tunisian dialects in Arabic Language, four main challenges are discussed including the very limited number of previous research conducted in this dialect, the lack of freely available resources for Sentiment Analysis, the absence of standard orthographies and code-switching with English or French.

[15,16] Compare the challenges between sentiment analyses of Tweets in the English Language versus Indian Regional Languages. Several challenges associated with the analysis of Twitter sentiments in Indian Regional Languages have been identified including Sarcasm Detection, Thwarted Expression, Negation Handling, Scarce resource language, Subjectivity detection and Domain Dependence.

In general, [17] conducted a review study of Sentiment Analysis for Non-English Languages. The main challenges are the availability of labeled non-English dataset, adequate techniques for pre-processing and lack of a lexicon dictionary consisting of a list of positive and negative sentiment words. Furthermore, [18] list the main challenges in the Sentiment Analysis field as Named entity recognition, Co-reference Resolution, Domain Dependency and Sentiment polarity detection.

Although Kurdish language is spoken by more than 40 million people around the world, it is still considered among the less resourced languages due to the unavailability of labeled corpus and inaccessibility of Natural Language Processing (NLP) tools [19]. Therefore, Above Challenges are among the main challenges that need to be addressed in Kurdish Language. Creating corpora and sentiment lexica is a major branch of related research for Kurdish Language. Moreover, having a large annotated corpus for sentiment analysis in Kurdish language yields a more accurate sentiment analysis.

3. The Steps of Processing Sentiment Analysis

Social media and websites present platforms for users to express opinions, thoughts, reviews, and sentiments towards a certain topic. As a result, a huge amount of data is gathered and need to be analysed to show the users' polarity (neutral, negative, or positive) about any entity. In this section, authors illustrate the processing steps of sentiment analysis practically,

and in each step the challenges of KSA are mentioned briefly while in section 4, the challenges are explained in details supported with examples. Figure 1 shows the workflow of sentiment analysis:

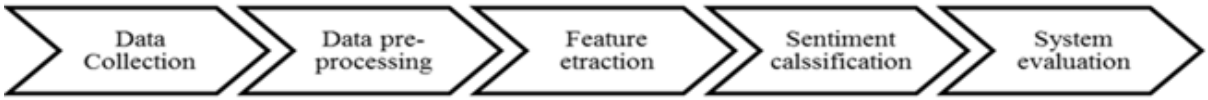

Figure 1 the workflow of sentiment analysis system

***3.1 Data collection*:** it plays a highly significant role in Kurdish Sentiment analysis. The quality and the quantity of the data effect on the classification performance of any used technique. A real challenge in Kurdish language is the diversity of writing the Kurdish text. As aforementioned users employ different scripts over the platforms social media and the websites which are different from the Modern Standard Kurdish (MSK).

***3.2 Data pre-processing*:** in this stage collected data is refined and non-textual content and the words that do not necessarily need to be analysed are removed before processing. For instance, Kurdish language like other non-English languages has no capitalization in its writing to specify names of an entity; this makes another challenge for Kurdish sentiment analysis. Moreover, other challenge is handling negations; Kurdish language has many words which are positive in writing while having a negative meaning.

3.3 Feature Extraction: this stage is very important in SA which extracts features from the analysed text. The features types are:

- **Part of speech:** it depends on the adjectives which reflect the opinion of individual. It is difficult to determine the adjectives from the diversity dialects/ scripts of Kurdish language.
- **Word frequencies and their presence:** this type relies on the frequencies of a word in the documents or only the presence of a word or N-gram words in the documents. These techniques are called TF and TF.IDF. The frequency count is implemented to illustrate the relative value of features [20].
- **Bag of words (BOW):** it is used to represent the text digitally. The challenging of applying this type in KSA is, all the vocabularies mentioned in the data collection need to be counted

with considering the diversity of dialects and the forms of the written Kurdish text. At the end, a huge amount of BOW with a big size of sparse matrix of zeros is generated which influences on the accuracy of the entire system. The following example shows how the sentence of (*I liked the movie*) is written in different dialects and forms in Kurdish language:

- فیلمەکەم بە دڵ بوو (Sorani Script using Kurdish letters) /felmækæm ba dl bo/
- *Filmakam ba dll bw* (Sorani Script using Latin letters)
- فیلمەکە یابدڵێمنبو (Kurmanji Script Kurdish letters) /felmækæ yab dlemnbu/
- *filmaka yab dlemnbu* (Kurmanji Script Latin letters)

3.4 Sentiment Classification: in this phase, the output of the system is generated and it should be under three categories (Positive, Negative, or Neutral).

3.5 System Evaluation: in this step system performance is evaluated through the metrics of recall, precision and f-measure.

4. Kurdish Language Sentiment Analysis Challenges

Nowadays, the majority of sentiment analysis studies have been done on the English language. However, despite Kurdish language being used by more than 40 million people around the world, the Sentiment analysis studies on this language are still in its early stages, due to the challenges and difficulties of Kurdish Language compared to the English language. In Kurdish sentiment analysis studies, researchers must consider the diversity of dialects and the all writing forms which make the practical part more challenging.

First of all, in terms of the way of writing, Kurdish language is written from right to left and it has different features of Orthography (norms of spelling, word breaks, punctuation, hyphenation, and emphasis) compared to other languages. In addition, the complex morphology and structure of the Kurdish language increases the level of challenges toward researchers. For instance, a word in Kurdish language might have different morphological aspects (derivation, inflection, and agglutination).

In addition, the diversity of Kurdish dialects and scripts create big challenges for data feature extraction. Kurdish language has two official scripts, Sorani and Kurmanji. Although there are some similarities in the alphabet of both scripts, there are huge differences in vocabularies and the structure of the sentences. The following **table 1** illustrates the most challenges are faced by researchers in Kurdish sentiment analysis. The first column of the table shows the

process steps in KSA, while the second column explain the challenges in each step, then more details about the challenges and the forth column shows examples of each case.

KSA processing steps	Challenges	Details	Example
Data collection	Scarcity of Supervised and unsupervised datasets and corpora for Sentiment Analysis	As the study of KSA is at very early stage, certainly there is Lack of existing datasets and corpus for sentiment analysis.	N/A
Data Preparation	Negation Handling	Negation Handling is one of the major challenges in Sentiment Analysis Especially when users express their negative feelings implicitly without using any negative keyword.	ئەوندە دڵ پڕم پێم خۆشە بگریم، کە فرمێسک و هەناسەش مەر هەمێکە / æwænda dl prm pem xôsh bgrem, kæ frmésk w hânâs mârhâmék/
Data Preparation	Code-Switching	Sometimes, users use more than one language to express their feelings such as using English Word or Arabic words inside Kurdish sentences	Unmuted someone and closed the app wtm bashkm aql bube shti be ma3na retweet naka launched twitter and the first thing I saw was her tetweet'y be ma3na muted again
Feature Extraction	Dialectical Kurdish rather than MSK	Using dialectical Kurdish rather than Modern Standard Kurdish (MSK), which is the formal written language. Mostly, Kurdish users on the social media platforms use colloquial Kurdish to express their sentiments. The colloquial Kurdish refers to the oral difference spoken among Kurdish people.	خیندەواریت نەبی ئەوهایە /xendâwâret nâbâ awhây/ (just like this if you are uneducated)
	Sarcasm Detection	Sarcasm is a special kind of sentiment which defines the inverse meaning of what people express in	ئەم شتە زۆر هەرزانە! / æm shtæ zôr hârz ânâ/

KSA processing steps	Challenges	Details	Example
		the text. Writing sarcastic sentences in Kurdish among social media users has gained popularity to avoid negative words. The use of strong positive words makes the sentence look positive but the overall meaning indicates negative due to the existence of sarcasm.	this product is very cheap!
	Script Diversity	Opinions are expressed in different dialects, especially Sorani and Kurmanji. Therefore, the feelings are expressed using different words.	The example is shown in section 3
Sentiment Classification	Lack of Sentiment Lexicon	Sentiment lexicons have not yet been created for Kurdish Language	N/A
Sentiment Classification	Kurdiz	Kurdish users when use Latin characters, they write numbers within the words which indicates the sound of a certain letter instead of writing the letter itself.	Good = bashi (Latin) = ba6i (Kurdiz) /bâshe/, here number 6 gives the sound /ʃ/

Table 1.Major Challenges Kurdish Sentiment Analysis processing steps

5. The proposed techniques

5.1 Hybrid approach

The most common techniques are used for sentiment classification are based-lexicon approach, machine learning approach, and hybrid approach [21]. Machine learning approach applies the used algorithms in text classification such as Support Vector Machine, Neural Network, Naïve Bayes, and etc. The Lexicon-based approach depends on sentiment lexicon, a collection of predefined words related to the sentiment and emotional of individual towards a certain subject. The lexicon-based approach is divided into two approaches dictionary-based and corpus-based approach. A collection of documents used to help, inference rules, learn and extract features is called Corpus. In addition, the hybrid approach is a combination of machine learning and lexical-based approaches. In [22,23,24,25] hybrid approach were applied to improve the efficiency in sentiment classification, and particularly a high efficiency was achieved in [26].

The proposed technique that might help to overcome some of the aforementioned challenges in Kurdish sentiment analysis is applying the hybrid approach of sentiment analysis; implementing corpus-based approach and support vector machine. Using corpus-based approach to build a comprehensive Kurdish lexicon that includes all the words with positive and negative polarity. In addition, improving the Kurdish corpus by adding lexicon-idioms which are phrases or expressions hold sentimental values. The idea is finding sentiment seed words, then searching for the synonyms of these words in all the forms of Kurdish dialects and scripts and inserting them to the dictionary.

In machine learning, the most important point is feature extraction and selection from the gathered data. These features are used to train the classifier. For feature extraction whether to rely on the frequency of the word occurrence in the documents or by depending on a term presence rather than its frequency. It has been achieved better results by using a word presence instead of word frequency [27]. In our proposed method, we recommend using a term presence for feature extraction in Kurdish sentiment analysis rather that term frequency tf.idf.

5.2 Lexicon-based approach for Kurdish sentiment analysis

In this paper, we recommend developing a new sentiment corpus which covers all the forms of written contemporary Kurdish language including formal and informal text. To the best of our knowledge, there is no Kurdish corpus for sentiment analysis that is why researches in Kurdish sentiment analysis are limited compare to other languages.

There have been numbers of works in developing sentiment corpus for non-English languages such as Arabic and Persian which are alphabetically very close to the Kurdish language. Opinion Corpus in Arabic [28] includes 500 movie reviews which are collected from Arabic websites and blogs. Each review was classified to positive and negative, then machine learning approaches (Support vector machine and Naïve Bayes) were applied to validate the results through comparing the performance of each algorithm.

The recent corpus was developed by [29] is SentiPers for Persian language. The corpus is composed of 26000 sentences and assigns a number from 1 to 5 to show the intensity positivity and negativity of an opinion of any given sentence.

The proposal is having a certain number of sentences (for example 20000 sentences) for each form of Kurdish script, Sorani, Kurmanji and Kurdish Latin script. The sentences should be collected from an official Kurdish websites and from social media platforms (facebook, tweeter,etc) to achieve the condition of formal and informal language. Then, assign number

within a specific range from 1 to 5 to specify the intensity of emotional orientation toward a certain entity for each sentence as in [29]. Finally, applying support vector machine on the corpus to validate the results of sentiment classification.

6. Conclusion And Future Work

This paper presented an overview of the main challenges and recent studies of Kurdish sentiment analysis. It is obvious that Dialect diversity, lack of language resources, distinct writing systems and Kurdiz are among the main challenges that need to be solved in order to enhance the performance of sentiment classification. Moreover, Building a large-scale corpus based on an accurate mapping for transliteration is highly recommended to overcome some of the challenges in KSA. In addition, a hybrid approach by combining machine learning approach and lexicon-based approach could be implemented to validate the approach performance on Kurdish language.

Reference

[1] Chandni, Chandra, N., Pahade, R., & Gupta, S. (2015). Sentiment Analysis and its Challenges. *International Journal of Engineering Research & Technology*, *4*(3). https://www.ijert.org/sentiment-analysis-and-its-challenges

[2] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

[3] Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics*, *9*(3), 483. https://doi.org/10.3390/electronics9030483

[4] Abdulla, S., & Hama, M. (2015). Sentiment Analyses for Kurdish Social Network Texts using Naive Bayes Classifier. *Journal of University of Human Development*, *1*, 393. https://doi.org/10.21928/juhd.v1n4y2015.pp393-397

[5] Esmaili, K. S. (2012). Challenges in Kurdish Text Processing. *ArXiv:1212.0074 [Cs]*. http://arxiv.org/abs/1212.0074

[6] Hassani, H., & Medjedovic, D. (2016). Automatic Kurdish Dialects Identification. In *Computer Science & Information Technology*(Vol.6).https://doi.org/10.5121/csit.2016.60307

[7] Walther, G., & Sagot, B. (2010). *Developing a Large-Scale Lexicon for a Less-Resourced Language: General Methodology and Preliminary Experiments on Sorani Kurdish*.

Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop). https://halshs.archives-ouvertes.fr/halshs-00751634

[8] Gautier, G. (1998). Building a Kurdish language corpus: An overview of the technical problems. *Proceedings of ICEMCO*.

[9] Oueslati, O., Cambria, E., HajHmida, M. B., & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. *Future Generation Computer Systems*, *112*, 408–430.

[10] Alsayat, A., & Elmitwally, N. (2020). A comprehensive study for Arabic sentiment analysis (challenges and applications). *Egyptian Informatics Journal*, *21*(1), 7–12.

[11] Alwakid, G., Osman, T., & Hughes-Roberts, T. (2017). Challenges in sentiment analysis for Arabic social networks. *Procedia Computer Science*, *117*, 89–100.

[12] Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134.

[13] Zahidi, Y., Younoussi, Y. E., & Yassine, A.-A. (2020). Arabic sentiment analysis problems and challenges. *2020 X International Conference on Virtual Campus (JICV)*, 1–4.

[14] Mdhaffar, S., Bougares, F., Esteve, Y., & Hadrich-Belguith, L. (2017). Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. *Third Arabic Natural Language Processing Workshop (WANLP)*, 55–61.

[15] Soman, S. J., Swaminathan, P., Anandan, R., & Kalaivani, K. (2018). A comparative review of the challenges encountered in sentiment analysis of Indian regional language tweets vs English language tweets. *International Journal of Engineering & Technology*, *7*(2.21), 319–322.

[16] Londhe, D. D., Kumari, A., & Emmanuel, M. (2021). Challenges in Multilingual and Mixed Script Sentiment Analysis. *2021 6th International Conference for Convergence in Technology (I2CT)*, 1–6.

[17] Djatmiko, F., Ferdiana, R., & Faris, M. (2019). A review of sentiment analysis for non-English language. *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, 448–451.

[18] Pandey, S. V., & Deorankar, A. V. (2019). A Study of Sentiment Analysis Task and It's Challenges. *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 1–5.

[19] Azad, R., Mohammed, B., Mahmud, R., Zrar, L., & Sdiqa, S. (2021). Fake News Detection in low-resourced languages "Kurdish language" using Machine learning

algorithms. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(6), 4219–4225.

[20] Y. Mejova and P. Srinivasan, "Exploring Feature Definition and Selection for Sentiment Classifiers," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, Art. no. 1, 2011.

[21] D. Maynard and A. Funk, "Automatic Detection of Political Opinions in Tweets," in *The Semantic Web: ESWC 2011 Workshops*, Berlin, Heidelberg, 2012, pp. 88–99. doi: 10.1007/978-3-642-25953-1_8.

[22] T. Lalji and S. Deshmukh, "Twitter sentiment analysis using hybrid approach," *International Research Journal of Engineering and Technology*, vol. 3, no. 6, pp. 2887–2890, 2016.

[23] I. Gupta and N. Joshi, "Enhanced twitter sentiment analysis using hybrid approach and by accounting local contextual semantic," *Journal of intelligent systems*, vol. 29, no. 1, pp. 1611–1625, 2020.

[24] M. E. Basiri and A. Kabiri, "Words Are Important: Improving Sentiment Analysis in the Persian Language by Lexicon Refining," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 17, no. 4, p. 26:1-26:18, May 2018, doi: 10.1145/3195633.

[25] B. Verma and R. S. Thakur, "Sentiment Analysis Using Lexicon and Machine Learning-Based Approaches: A Survey," in *Proceedings of International Conference on Recent Advancement on Computer and Communication*, Singapore, 2018, pp. 441–447. doi: 10.1007/978-981-10-8198-9_46.

[26] V. Nandi and S. Agrawal, "Political sentiment analysis using hybrid approach," *International Research Journal of Engineering and Technology*, vol. 3, no. 5, pp. 1621–1627, 2016.

[27] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *INR*, vol. 2, no. 1–2, pp. 1–135, Jul. 2008, doi: 10.1561/1500000011.

[28] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega, "OCA: Opinion corpus for Arabic," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 10, pp. 2045–2054, 2011, doi: 10.1002/asi.21598.

[29] P. Hosseini, A. A. Ramaki, H. Maleki, M. Anvari, and S. A. Mirroshandel, "SentiPers: A Sentiment Analysis Corpus for Persian." arXiv, Jan. 01, 2021. doi: 10.48550/arXiv.1801.07737.