



Article

AraSenCorpus: A Semi-Supervised Approach for Sentiment Annotation of a Large Arabic Text Corpus

Ali Al-Laith ^{1,*} , Muhammad Shahbaz ¹, Hind F. Alaskar ²  and Asim Rehmat ¹

¹ Computer Science Department, University of Engineering and Technology, Lahore 54890, Pakistan; m.shahbaz@uet.edu.pk (M.S.); asimrehmat@uet.edu.pk (A.R.)

² Artificial Intelligence and Data Analytics Laboratory, Prince Sultan University, Riyadh 11586, Saudi Arabia; halaskar@psu.edu.sa

* Correspondence: ali.allaith@kics.edu.pk

Abstract: At a time when research in the field of sentiment analysis tends to study advanced topics in languages, such as English, other languages such as Arabic still suffer from basic problems and challenges, most notably the availability of large corpora. Furthermore, manual annotation is time-consuming and difficult when the corpus is too large. This paper presents a semi-supervised self-learning technique, to extend an Arabic sentiment annotated corpus with unlabeled data, named AraSenCorpus. We use a neural network to train a set of models on a manually labeled dataset containing 15,000 tweets. We used these models to extend the corpus to a large Arabic sentiment corpus called “AraSenCorpus”. AraSenCorpus contains 4.5 million tweets and covers both modern standard Arabic and some of the Arabic dialects. The long-short term memory (LSTM) deep learning classifier is used to train and test the final corpus. We evaluate our proposed framework on two external benchmark datasets to ensure the improvement of the Arabic sentiment classification. The experimental results show that our corpus outperforms the existing state-of-the-art systems.



Citation: Al-Laith, A.; Shahbaz, M.; Alaskar, H.F.; Rehmat, A.

AraSenCorpus: A Semi-Supervised Approach for Sentiment Annotation of a Large Arabic Text Corpus. *Appl. Sci.* **2021**, *11*, 2434. <https://doi.org/10.3390/app11052434>

Academic Editor: Carlos A. Iglesias

Received: 1 February 2021

Accepted: 26 February 2021

Published: 9 March 2021

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: corpus annotation; Arabic sentiment analysis; semi-supervised learning; self-learning; neural networks; deep learning

1. Introduction

Several tasks in natural language processing require annotated corpora for training and evaluation methods and comparing the different systems [1]. The process of manual annotation of corpora is usually costly and becomes prohibitive when scaled to a larger dataset [2]. For popular tasks on natural language processing, such as sentiment analysis, we can find widely used corpora that serve as baselines for approaches and methods proposed for the sentiment analysis task. For example, two datasets are the SemEval 2017 [3] and Arabic Sentiment Tweets Dataset (ASTD) [4], which are used for evaluating state-of-the-art models due to the reliability of the annotation, and both corpora contain a large number of annotated documents.

In the literature, there are several corpora for the task of Arabic sentiment analysis, but the high costs associated with manual annotation limit these resources to be either small or obtained through entirely automatic methods such as user rates or Arabic sentiment lexicons. Furthermore, the data presented in these corpora are outdated, incomplete, or small.

In this paper, we introduce AraSenCorpus, a semi-supervised framework to annotate a large Arabic text corpus using a small portion of manually annotated tweets (15,000 tweets) and extending it from a large set of unlabeled tweets (34.7 million tweets) to reduce human effort in annotation and providing a middle ground between manual and automatic labeling of a large dataset. We used the FastText neural network [5], and along-short term memory (LSTM) deep learning classifier to expand the manually annotated corpus and ensure the quality of the newly created corpus, respectively. The outcome of the developed

corpus is tested using a set of Arabic benchmark datasets. The FastText algorithm is an open-source NLP library developed by Facebook AI. It is a fast and excellent tool to build NLP models and generate live predictions [6]. LSTM with word embeddings is used to perform the sentiment classification, as this classifier outperforms the traditional techniques in text classification [7]. The classifier performed well with embeddings, especially when dealing with the sentiment classification of Arabic dialects [8].

The rest of the paper is organized as follows: in Section 2, we present a literature survey of Arabic sentiment corpus construction and sentiment analysis and discuss their approaches and performance. In Section 3, we present our methodology for the corpus generation process and show in detail the data collection and sentiment annotation, and statistical information about the corpus. In Section 4, we present the experimental results and analysis of the manual and automatic annotation on our corpus. In Section 5, we conclude with a discussion of our work and future plans, and then we sum up our contributions and present the conclusion in Section 6.

2. Related Work

The creation of corpora for sentiment analysis was largely addressed in the related work. Several types of research have been made using automatic labeling techniques due to the challenges involved in manual annotation such as training annotators, measuring the inter-annotator agreement, writing the annotation guidelines, and developing the annotation interfaces [9]. In the following two sections, we will present literature studies about sentiment annotation and classification approaches, with the main focus being on Arabic sentiment analysis studies.

2.1. Sentiment Annotation

Arabic text can be classified into three categories: (a) classical Arabic, the version in which the Quran (the holy book of Islam) is written in, (b) modern standard Arabic (MSA), which is used in all Arabic countries for official and formal purposes such as newspapers, schools, and universities, and (c) dialectical Arabic (DA), which is used in everyday life among the people in different regions. For sentiment analysis, most of the existing textual corpora are either MSA or DA. The sentiment annotation of Arabic text can be classified into three approaches: (1) automatic annotation, (2) semi-automatic annotation, and (3) manual annotation. These three annotation approaches are summarized in Figure 1. Several studies have been conducted using automatic techniques for corpus construction and annotation. For this purpose, three main techniques have been used: (1) automatic annotation based on rating reviews [10–14], (2) sentiment lexicon [15–18], and (3) external application programming interfaces (APIs) [19]. In the context of sentiment annotation based on rating reviews, a Large-Scale Arabic Book Reviews (LABR) corpus was proposed in [10] for a specific domain. This corpus is a collection of book reviews containing 63,257 book reviews that had been written by readers. Each review is on a scale from 1 to 5 based on the user's rating of books. The authors considered reviews with a score of 1 or 2 as negative, 4 or 5 as positive, and 3 as neutral. Another sentiment corpus is called "BRAD 1.0", a large sentiment corpus that was created in [12] and motivated by the LABR corpus in [10]. This corpus consists of 510,600 book reviews. It was collected from the goodreader.com website. The user's rating of each book review on a scale of 1–5 was used to annotate the corpus reviews by following [10]. Along with modern standard Arabic, the corpus also includes some dialects such as Egyptian dialect content. BRAD 2.0 is an extension of BRAD 1.0. It contains more than 200K extra reviews to account for more Arabic dialects [11]. To classify the reviews, they followed the same procedures as in [10,12]. User ratings are used to annotate each review in the corpus into positive, negative, or neutral classes. Similarly, among the presented corpora in [12,14], HARD is the most recent sentiment corpus for hotel reviews [14]. This corpus consists of more than 370,000 reviews. It was collected from the booking.com website and automatically annotated based on the reviews rating, but on a scale of 1–10. Each review is annotated

as positive, negative, or neutral based on the user's rating of the review. Similar work was done in [11]. The authors annotated a collection of reviews from different domains such as hotels, restaurants, movies, and product reviews. The reviews were extracted from different websites. The annotated corpus contains 33,000 reviews.

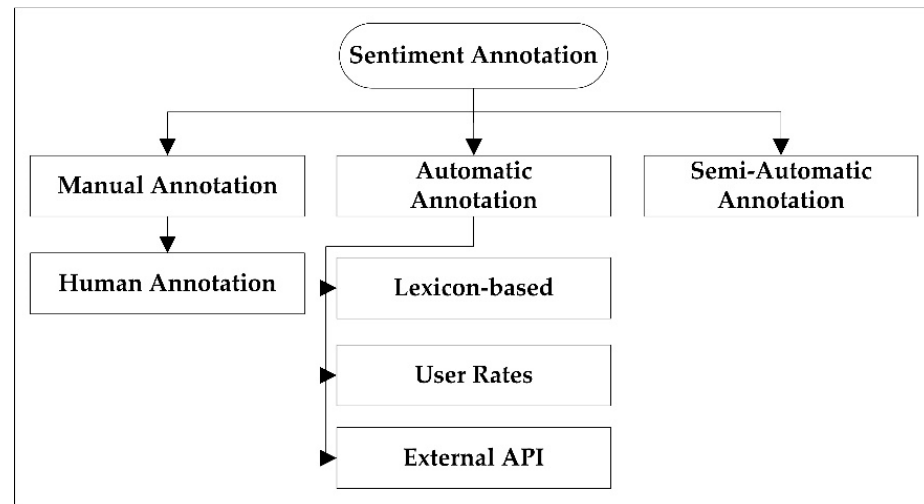


Figure 1. Sentiment annotation approaches.

The second approach of the automatic annotation is the use of sentiment lexicons to automatically annotate the textual corpora. In [15], an Algerian sentiment lexicon was automatically constructed by translating English sentiment lexicons. Based on the created sentiment lexicon, a corpus containing 8000 messages, which were written in Arabic and Arabizi (writing Arabic using English characters), was automatically annotated into positive and negative classes. Another sentiment corpus was proposed in [16]. It contains 151,548 tweets in modern standard Arabic (MSA) and Egyptian dialects. The tweets were classified into positive and negative classes. They relied on the corpus proposed in [10] by manually extracting and annotating a list of 4404 phrases that are commonly used in expressing sentiment, to be used in the annotation of their corpus. The annotation was performed based on the manually annotated phrases and the frequency of positive and negative words appearing in the text. Another work was done in [20] by collecting 10 million tweets from Twitter and using a list of positive and negative Arabic words as search keywords. The annotation was performed automatically by considering tweets containing positive search keywords as positive tweets and similarly tweets containing negative search keywords as negative tweets. They used the Twitter API to collect tweets during June and July 2017. Before classifying the tweets, they performed further text preprocessing steps to clean the collected dataset and obtain better results. The use of distant supervision is another approach for automatic annotation. TEAD is a dataset for Arabic sentiment analysis proposed in [17]. The dataset contains more than 6M tweets that were collected from Twitter using a list of emojis as search keywords. After filtering the tweets, automatic annotation grounded in a lexicon-based approach was performed for sentiment analysis (Ar-SeLn sentiment lexicon). The annotation method in this research was evaluated manually by extracting 1000 tweets in each class.

The third approach of the automatic annotation is the use of external APIs such as a tool called "AYLIE", which was used to annotate a sentiment corpus for the Sudani dialect [19]. The corpus was collected from Twitter. It contains 5456 tweets and automatically classifies them into three classes: positive, negative, and neutral.

The semi-automatic approach is the second type of corpus annotation. AraSenTi-Tweet is a sentiment corpus that contains 17,573 tweets [21]. The corpus text is written in the Saudi dialect. A sentiment lexicon was used to collect tweets that contain such words. After preprocessing and cleaning the tweets, three annotators were employed to review

the constructed corpus. Similar work was proposed in [22] for the Saudi dialect as well. They used a sentiment lexicon to collect the corpus text from Twitter. The extracted tweets were filtered and manually annotated into two different classes, positive and negative. The resulting corpus contains 4000 tweets. The semi-supervised annotation approach has been applied for other languages, such as the Brazilian Portuguese language. In [23], the authors extended a small sentiment corpus, which was annotated manually, to annotate a large unlabeled corpus. They used only one classifier to predict the classes of the unlabeled documents and added those documents which have a confidence value above a predefined threshold. Another large unlabeled dataset of tweets was presented for the English language in [24]. A total of 384 million tweets were collected in the entirety of the year 2015. They used two semi-supervised learning approaches, self-learning and co-training, to annotate a huge collection of tweets.

The last approach for corpus annotation is the manual annotation approach. ASTD is an Arabic sentiment corpus that contains 10,000 tweets which were classified as positive, negative, mixed, and objective [19]. They used Amazon Mechanical Turk [25] to annotate the tweets in the dataset. The crowdsourcing technique is a method for manual annotation [26,27]. In [26], crowdsourcing was used to classify the tweets into two classes, positive or negative. This corpus contains 32,063 tweets written in the Saudi dialect. Human annotation is an approach related to manual annotation. ArSentD-LEV is a Levantine dialect sentiment corpus that was proposed in [27]. The corpus contains 4000 tweets. It was annotated manually with different annotations including the overall sentiment of the tweet. The corpus was annotated with five-point scale classes: very positive, positive, neutral, negative, and very negative. The annotation process was manually carried out via crowdsourcing using the CrowdFlower platform [28]. In [29], a sentiment corpus SANA for the Algerian dialect was collected from the web. The corpus contains 513 comments which are classified into positive or negative classes. Two Algerian Arabic native speakers were employed to annotate the corpus. In [30], a sentiment and emotion corpus was constructed from Twitter. Three annotators engaged in the classification process where they labeled each tweet according to its sentiment polarity (e.g., positive, negative, and neutral). The corpus contains 5400 tweets. Another dialect sentiment corpus was proposed in [31] for Jordanian dialect tweets. The corpus was manually annotated into three classes—positive, negative, and neutral—by Arab Jordanian students. It contains 1000 tweets. A customer review corpus called “MASC” was collected from multiple websites and social media platforms [32]. It was manually annotated by two native speakers into positive and negative reviews. The corpus covered 15 different domains such as art and culture, bakeries and goodies, cafes, fashion, financial services, hotels, restaurants, etc. The corpus contains 8860 reviews. MSAC is a multi-domain sentiment corpus that covers different domains, such as sport, social issues, and politics [33]. The corpus contains 2000 tweets that were manually annotated into two classes: positive and negative. A Tunisian dialect sentiment corpus called “TSAC” was presented in [34]. The corpus contains 17,000 user comments from Facebook that were collected and annotated manually into two classes, positive and negative. The proposed corpus is a multi-domain corpus consisting of vocabulary from the education, social, and political domains. AWATIF, a multi-genre sentiment corpus, was introduced in [35]. The corpus contains 10,723 Arabic sentences retrieved from three resources: the Penn Arabic Treebank, web forums, and Wikipedia. The corpus was manually annotated as objective and subjective (both positive and negative). Another two Arabic sentiment corpora from Twitter were introduced in [36,37]. In [36], the corpus contains 2300 tweets that were manually annotated, while in [37], the corpus contains 2000 tweets. The tweets were classified as positive and negative by native annotators.

To summarize, several Arabic sentiment corpora have been developed for the Arabic sentiment classification task. Table 1 shows most of the existing Arabic sentiment corpora. The annotation of these corpora was conducted using three approaches: manual, automatic, and semi-automatic. The data of such corpora were collected either from social media platforms, customer reviews, or comments. The corpora based on social media texts are

mostly generated from Twitter. It is observed that the size of automatically annotated corpora is larger than those which were manually or semi-automatically annotated. The semi-automatic annotation was performed using sentiment lexicon terms to extract data and manual annotation to annotate the extracted data. This paper contributes a semi-automatic Arabic sentiment corpus which contains 34.7 million tweets and spans 14 years. A semi-automatic approach was applied to annotate the corpus using a self-training approach which is, to our knowledge, not used in any of the existing Arabic sentiment corpora. This approach helps in building large-scale labeled datasets that span large periods of time, as reported in [24].

Table 1. Existing Arabic sentiment annotated corpora (MSA: Modern Standard Arabic and DA: Dialectical Arabic).

Study Dataset	Year Size	Domain	Annotation Method	Type
[10] Labr	2013 63,257	Book reviews	Automatic	DA
[11] HTL, RES, MOV, PROD	2015 33,000	Movies, hotels, restaurants, and products reviews	Automatic	MSA/DA
[12] BRAD 1.0	2016 510,600	Book reviews	Automatic	MSA
[13] BRAD 2.0	2018 692,586	Book reviews	Automatic	MSA/DA
[14] HARD	2017 490,587	Reviews	Automatic	MSA/DA
[15] SentiALG	2018 8000		Automatic	DA
[16]	2018 151,000	Tweets	Automatic	DA
[19] SSA-SDA	2019 5456	Politics	Automatic	DA
[17] TEAD	2018 6M	Tweets	Automatic	MSA/DA
[21] AraSenTi-Tweet	2017 17,573	Tweets	Semi-automatic	MSA/DA
[22]	2018 4000	Tweets	Semi-automatic	MSA/DA
[4] ASTD	2015 10,000	Tweets	Manual	MSA/DA
[26] SDCT	2019 32,063	Tweets	Manual	DA
[27] ArSentD-LEV	2019 4000	Tweets	Manual	DA
[29] SANA	2019 513	Comments	Manual	DA
[30]	2018 5400	Tweets	Manual	DA
[31]	2019 1000	Tweets	Manual	DA
[32] MASC	2018 8860	Reviews	Manual	MSA/DA

Table 1. *Cont.*

Study Dataset	Year Size	Domain	Annotation Method	Type
[33] MSAC	2019 2000	Reviews	Manual	DA
[34] TSAC	2017 17,000	Facebook Comments	Manual	DA
[35] AWATIF	2012 10,723	Penn Arabic Treebank, web forums, and Wikipedia	Manual	MSA/DA
[36]	2013 2300	Tweets	Manual	MSA/DA
[37] ArTwitter	2013 2000	Tweets	Manual	DA

2.2. Sentiment Classification

The sentiment classification is the task of classifying a document into positive, negative, or neutral classes [38]. There are different sentiment classification approaches and tools used for the sentiment classification task on the document level that considers the whole document as a basic information unit. In this part, we will describe the approaches, features, and techniques used along with types of testing and validation of the existing sentiment corpora with a focus on the Arabic sentiment corpora.

From the revised literature, it was proved that machine learning approaches (e.g., support vector machine, naïve Bayes, and logistic regression) are the classifiers used for sentiment classifications, as shown in Table 2. These classifiers are suitable for the case of Twitter data [4,15,21,24,25,28,33,34] and YouTube data [39].

Table 2. Sentiment classification approaches: machine/deep learning classifiers commonly used in performing Arabic sentiment classification.

Classification	Approaches	Features/Techniques	Studies
Machine Learning	Support Vector Machines (SVM), Decision Tree (DT), Naïve Bayes (NB), and Multinomial Naïve Bayes (MNB)	Term Frequency—Inverse Document Frequency (TF-IDF), N-grams, Count Vector	[4,10–13,15,16,19,22,26,27,29,32–34,40]
Deep Learning	Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (B-LSTM), and Convolutional Neural Network (CNN)	Continuous Bag of Words (CBOW)	[16,19,24,30]
Lexicon-based		Lexicon Terms	[10,13,30]

Many studies used deep learning models to recognize and understand data such as text, audio, and images. Long short-term memory (LSTM) and convolutional neural network (CNN) deep learning classifiers were used in [17]. They used both classifiers with a huge amount of training data to achieve better results. In [26], Bi-LSTM and LSTM achieved 94% and 92% accuracy for a dataset of 32,063 tweets, respectively. CNN and LSTM are two deep learning models that were used in [33]. The models achieved F-measures of 95% and 93%, respectively, by using a word2vec pre-trained embedding as input features for both models. An online system for Arabic sentiment analysis called “Mazakak” was presented in [41]. The system was developed based on a CNN followed by an LSTM deep learning model. It achieved state-of-the-art results on two benchmark datasets, SemEval

2017 and ASTD. The proposed approaches represent documents on counting, N-grams, TF-IDF, and word embeddings features.

Lexicon-based approaches were also used to perform sentiment analysis. In [42], an Arabic lexicon was constructed and combined with a named entity recognition (NER) system to show the importance of including NER in the process of sentiment analysis. The experimental results, on different Arabic sentiment corpora, obtained by using NER outperformed the results obtained without using NER. NileULex is an Arabic sentiment lexicon for modern standard Arabic and Egyptian dialect [43]. The lexicon was used over two manually annotated datasets to ensure that the developed lexicon is useful for Arabic sentiment analysis. The authors reported that the use of their lexicon, which is manually constructed and annotated, gives better results in comparison with the translated or automatically constructed lexicons. In [44], the sentiment polarity was calculated by looking for the polarity of each term of the message in the constructed lexicon. They performed several preprocessing steps including tokenization, normalization, repeat letters and stop words removal, light stemming, and negation and intensification handling. Two Arabic sentiment corpora were used to perform classification experiments, and the best accuracy result was 70%.

3. Methodology

3.1. Data Collection Process

This section describes the process of data collection and the challenges we faced during the data gathering. It also presents our data collection methodology, representatives of our collected data, characteristics, and distribution of the sentiment corpus, and the potential applications of AraSenCorpus.

3.1.1. Challenges in Data Collection from Twitter

On a monthly basis, Twitter serves around 330 million active users [45], and 500 million tweets per day. While Twitter is public and tweets are viewable and searchable by anyone around the world, there are specific challenges to collect Twitter data. For example, the standard API only allows for retrieving tweets from up to 7 days ago, scrapping a limited number of tweets per 15min window. To overcome this problem, we developed a python script that can retrieve tweets over a long period.

3.1.2. Challenges in Using Twitter as a Data Source

Using Twitter as a data source in academic research leads to some challenges that may be faced, such as:

- Ethical issues: Reproducing tweets in an academic publication has to be handled with care, especially concerning tweets related to sensitive topics. In our research, our topic targets sentiment analysis for Arabic text where the extracted tweets are not sensitive.
- Legal issues: Under Twitter's API Terms of Service, it is prohibited to share Twitter datasets. Our corpus will be available for the research community by sharing the IDs of tweets only, which can be used by other researchers to obtain the tweets, along with their sentiment labels.
- Retrieving datasets: Using certain keywords may not retrieve all of the tweets related to a topic. In our research, it is also a challenge to build a sentiment corpus by searching for a limited number of keywords in Arabic, while there are many Arabic dialects as well. We overcome this problem by collecting sentiment terms/phrases from multiple Arabic sentiment lexicons in modern standard Arabic and Arabic dialects.
- Cost: Twitter data cost a lot of money when they are obtained from a licensed reseller of Twitter data. It is also difficult to obtain Twitter data using the free API. However, our developed system can obtain a large amount of data.
- Spam: There are large numbers of tweets on Twitter that can attract a large amount of spam. In our research, we found a lot of tweets that contained sexual expressions that had to be excluded from the sentiment corpus.

3.1.3. Data Collection Methodology

To build the AraSenCorpus corpus, we selected Twitter as the data source for data collection. Twitter is a rich platform to learn about people's opinions and sentiments on different topics as they can share their opinions and thoughts. Twitter is considered a rich resource for sentimental text, containing views on many different topics: social issues, politics, business, economics, etc. A list of sentiment terms/phrases was obtained from Arabic sentiment lexicons in modern standard Arabic and Arabic dialects, as shown in Table 4. The list was verified and considered to represent search keywords to obtain tweets from Twitter social media platform. The selection of data sources and search query keywords is a very important part of the study.

For data collection, the list of sentiment terms/phrases helped us to ensure that the collected tweets are diversified and they are representative of many sentiment tweets from different topics. The collection system used these sentiment terms/phrases and retrieved tweets spanning from 2007 to 2020. The statistics about the collected tweets are shown in Table 3.

Table 3. Twitter dataset statistics.

Title	Number
Tokens	479,661,838
Unique Tokens	9,281,106
Average Words per Tweet	13.8
Users	7,649,717
Total Tweets	34,706,737

3.1.4. Representative Consideration

To collect a large corpus while ensuring that it covered many sentiment terms and phrases, five sentiment lexicons were used to extract the corpus tweets. The lexicons used in this research were: (1) 230 Arabic words(modern standard Arabic) [46], (2) Large Arabic Resources for Sentiment Analysis(modern standard Arabic) [47], (3) MorLex lexicon (Modern Standard Arabic and Egyptian Dialect), (4) NileULexlexicon (modern standard Arabic and Egyptian dialect) [48], and (5) Arabic senti-lexicon(modern standard Arabic) [49]. Details about these lexicons are shown in Table 4.

Table 4. Five Arabic sentiment lexicons: terms/phrases used as search keywords to extract tweets from Twitter.

#	Lexicon	# of Terms/Phrases	Type	Study
1	230 Arabic words	230	MSA	[50]
2	Large Arabic Resources for Sentiment Analysis	1913	MSA/DA	[11]
3	MorLex	10,761	MSA/DA	[51]
4	NileULex	5953	MSA/DA	[43]
5	senti-lexicon	3880	MSA/DA	[32]

3.1.5. Corpus Cleaning and Preprocessing

The text of tweets is known to be noisy and should be cleaned and preprocessed before performing sentiment classification in order to get better results. While collecting tweets, those that contained URLs, hashtags, mentions, or media were already cleaned before adding them to the corpus. The tweets were processed for text tokenization and normalization. Normalization is the process of unifying the shape of some Arabic letters that have different shapes. For example, the Arabic letters (و, ي, ة, أ) are normalized to convert multiple shapes of the letter to one shape, the different forms of "Alef" (أ, إ, آ) are

converted into (ل), the different forms of “Ya’a” (ي,ى) are converted into (ي), the letter “Ta’a” (ة) is converted into (o), and the letters ع and ؤ are converted to (e). Non-Arabic letters such as (!, -, &, *) are removed by iterating all the tweet words to remove the noise from the text.

Repeated characters add noise and influence the mining process, which makes it very difficult. For example, a word may be written like “رأع” instead of “رائع”, which means “wonderful”. To resolve this, we returned the word to its correct and right syntax by removing the extra repeated characters. Diacritics are rarely used on social media websites. These diacritics, in most cases, are used to add decorations to the text. While preprocessing, we removed all diacritics from the tweets. Furthermore, the duplicated tweets were already excluded from the corpus in the collection phase by looking for the tweet’s ID and removing the tweet if it already exists in the corpus.

3.1.6. Corpus Characteristics and Potential Applications

The collected corpus contains more than 34.7 million tweets. We used more than 22,000 terms/phrases to collect tweets from Twitter. These terms/phrases were extracted and manually verified from 5 different Arabic sentiment lexicons. The total number of tokens in the corpus exceeds 479 million tokens, while the unique number of tokens is more than 9 million tokens. More than 7.6 million users participated in the collected corpus. Table 3 shows statistics about the collected tweets.

Zipf’s law states that the frequency of word tokens in a large corpus is inversely proportional to the rank [52]. The law states that if f is the frequency of a word in the corpus and r is the rank, then:

$$f = \frac{k}{r} \quad (1)$$

where k is a constant for the corpus. We follow [53] to verify Zipf’s law by calculating the log of the frequency of word tokens and their ranking in our corpus. Figure 2 shows the Zipf’s law curve of unigrams frequencies and their ranking. The curve ensures that our corpus does not have anomalous biases.

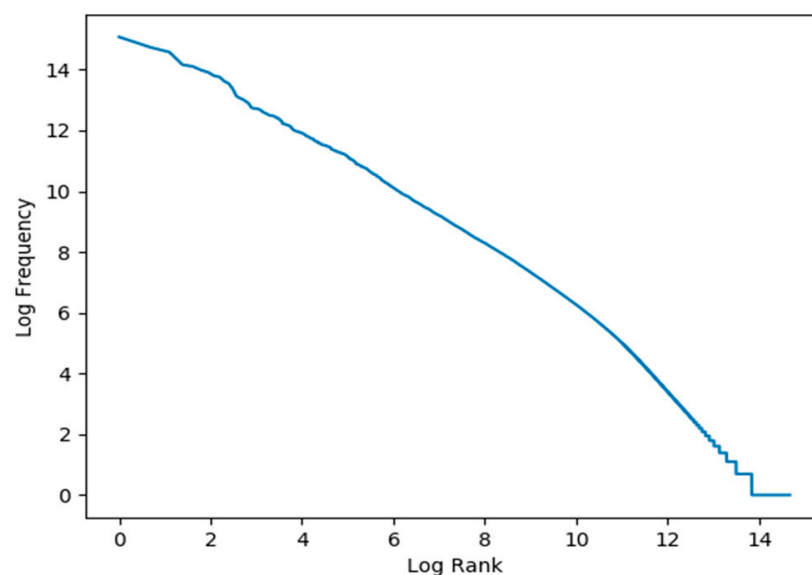


Figure 2. Zipf’s curve for AraSenCorpus unigrams.

Most of the proposed Twitter sentiment corpora in the literature were collected within a few months [4,17,21,27], or less than two years [30]. This does not ensure the coverage of multiple topics in social media. Our corpus contains tweets that were collected over 14 years, starting from 2007 and ending in 2020. Table 5 shows distribution of the collected

tweets. The highest number of tweets in the corpus was collected in the year 2020, while a low number of tweets were collected in the year 2007. This indicates the fact that the number of users of social media has increased over time, which causes the increase in the amount of data on social media. Another reason is the use of the Twitter API, which retrieves tweets from the last 7 days, and that the collection started during the year 2020.

Table 5. Distribution of tweets in the corpus.

Year	# of Tweets	Year	# of Tweets
2007	116	2014	385,841
2008	756	2015	824,102
2009	65,287	2016	2,238,905
2010	730,772	2017	1,395,090
2011	291,451	2018	3,514,951
2012	336,933	2019	5,382,799
2013	214,461	2020	19,325,273

We believe that our proposed corpus can be used in different applications. In particular, the proposed sentiment corpus can be very helpful in (1) enhancing the existing Arabic sentiment analysis approaches and (2) building domain-specific pre-train sentiment models (such as word2vec, Glove, and BERT) for Arabic sentiment analysis.

3.2. Corpus Framework

3.2.1. Manually Annotated Dataset

Two Arabic native speakers were requested to annotate our manual corpus. Set of annotation guidelines were given to the annotators to provide the best degree of contingency in the obtained results.

Annotation Guidelines: The annotation guidelines were defined to label tweets in our corpus. We first surveyed the existing work on annotation guidelines to define the baseline guidelines. Second, we improved these guidelines. Two annotators were asked to independently annotate 3000 tweets under the provided guidelines. The third annotator was employed to resolve ambiguous cases if they were found during the annotation process. The ambiguous tweets included tweets with mixed sentiments in a single tweet and tweets in special cases such as sarcastic tweets. Three main aspects were formulated before performing the annotation.

1. What to annotate: Tweets that bear a positive or negative sentiment and tweets that do not bear any positivity or negativity (neutral tweets).
2. What not to annotate: tweets containing both positive and negative sentiments.
3. How to handle special cases such as negations, sarcasm, or quotations.

The following are the guidelines given to the annotators for manual annotation:

1. Tweets bearing positive terms/phrases: for instance, “رائع” (wonderful) and “بسيطة ومفيدة” (simple and useful).
2. Tweets bearing negative terms/phrases: for instance, “سيئ” and “تجربة مريرة” (difficult experience).
3. Positive or negative situations or events, for example, “تسبب فيروس كورونا” “بجسائر فادحة لمعظم دول العالم” (corona virus caused heavy losses to most of the world countries). This tweet will be marked as negative since it has two negative terms “جسائر” (losses) and “فادحة” (fatal).
4. All objective tweets that do not contain any sentiment terms/phrases will be considered neutral and marked as “neutral”.

5. Tweets containing both positive and negative terms/phrases with the same intensity of positive and negative terms/phrases will be marked as “mixed” and will not be considered in this research.
6. Tweets containing negations before positive or negative terms/phrases will flip the polarity of sentiment, for example “غير مريحة” (uncomfortable) is a phrase with negation and a positive term. Tweets containing such phrases should be marked as negative tweets.

A small .Net application was designed to facilitate the annotation process, as shown in Figure 3.

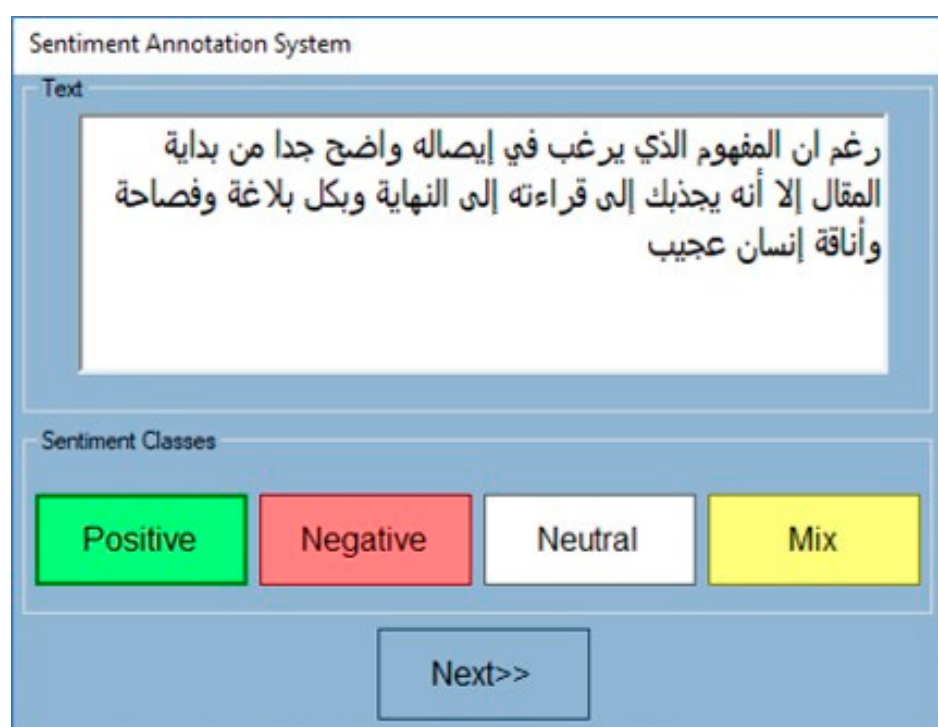


Figure 3. Sentiment annotation interface. The English translation of the tweet is: *Although the concept that he wishes to convey is very clear from the beginning of the article, it will draw you to read it to the end, with all eloquence, eloquence and elegance, a wonderful person.*

Inter-annotator Agreement: To verify the completeness of the annotation guidelines, we gave 3000 tweets to two different annotators and asked them to annotate the given dataset. They independently annotated these tweets into “positive”, “negative”, and “neutral” classes based on the given annotation guidelines and using the annotation application as shown in Figure 1. The inter-annotator agreement was 0.93 (using Kohen’s kappa coefficient), which is considered to be a perfect value. The remaining 12,000 tweets were divided equally between both annotators to speed up the annotation process. We included tweets that belonged to the positive, negative, and neutral classes only.

3.2.2. Semi-Supervised Annotated Corpus

Initially, we used the self-learning technique by following [54] to expand the manually annotated dataset (15,000 tweets). As shown in Figure 4, we trained three classifiers on the manually annotated dataset. These classifiers were built using the FastText neural network algorithm. This algorithm was selected as it achieves similar results to the machine/deep learning classifiers while training a lot faster, as reported in the initial paper [6]. Using this algorithm, we can train and test a model, predict sentiment classes of tweets, and predict the probability of tweets towards sentiment classes.

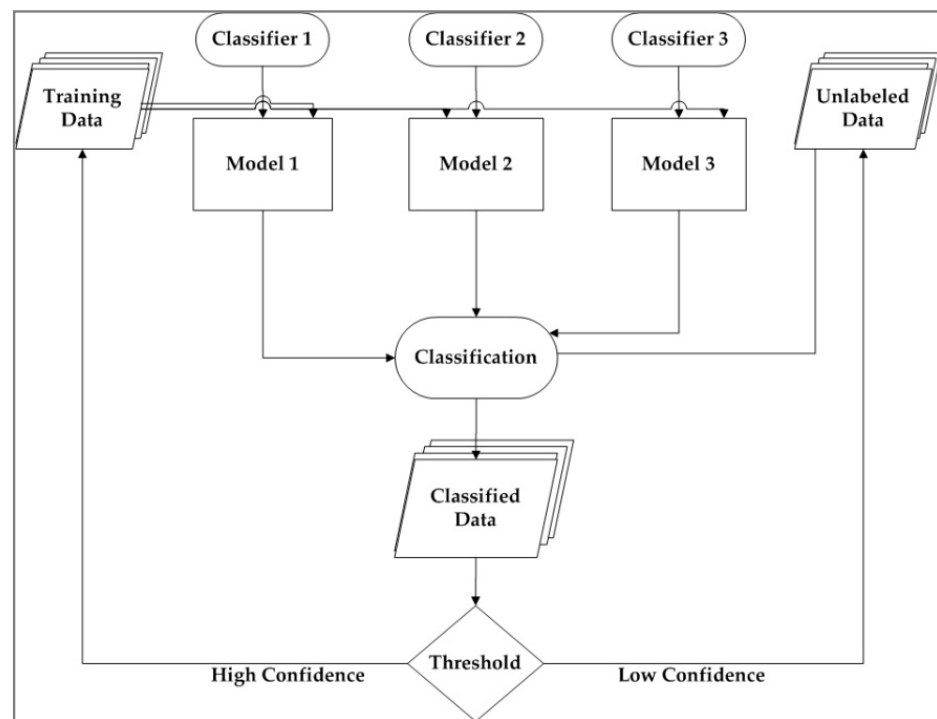


Figure 4. AraSenCorpus annotation framework.

The following are different feature approaches used for handling the data:

1. Classifier#1: Learning rate (LR) = 0.5, epoch = 10, wordNgrams = 1, and dimension = 100.
2. Classifier#2: Learning rate (LR) = 0.5, epoch = 10, wordNgrams = 2, and dimension = 100.
3. Classifier#3: Learning rate (LR) = 0.5, epoch = 10, wordNgrams = 3, and dimension = 100.

Each tweet from the unlabeled tweets passed through three classifiers to be annotated with the sentiment class and the probability of the assigned class. If a tweet had a confidence value greater than or equal to a threshold (e.g., 90%) in all classifiers, it was added to the training data and removed from the unlabeled data. We performed this task 3 times to increase the size of the labeled data. Moreover, we ensured the quality of the labeled data using a set of benchmark datasets after every iteration.

AraSenCorpus differs from other semi-supervised approaches in the literature. Usually, the authors use a sentiment lexicon to annotate unlabeled data and revise a sample of the annotated data manually. Since this approach needs more time to revise the annotation, we proposed a self-learning approach to automate the annotation and reduce human effort, as shown in Algorithm 1. The intuition behind AraSenCorpus is that the manual annotation is necessary for subjective tasks such as sentiment analysis and should be part of the process. This is in addition to the fact that the iterative addition of new tweets provides new information for the classifiers, thus resulting in better classification models for labeling the remaining unlabeled tweets.

Algorithm 1: Semi-Supervised (Self-Learning) Annotation Algorithm.

```

Data: (LabeledData (15K Tweets), UnlabeledData (34.7M Tweets))

Result: NewLabeledData
TrainingSet = LabeledData;
UnlabeledData = UnlabeledData;
Iterations = 3;
ThresholdValue = 0.9;
while( $I \leq \text{Iterations}$ )do
  //training 3 different fastText models on TrainingSet
  Model1 = TrainClassifier1(TrainingSet);
  Model2 = TrainClassifier2(TrainingSet);
  Model3 = TrainClassifier3(TrainingSet);
  while( $t \leq \text{UnlabeledData.size}()$ )do
    //predict most likely sentiment classes of t from model1, model2, and model3
    S1 = Model1.predict(UnlabeledData(t));
    S2 = Model2.predict(UnlabeledData(t));
    S3 = Model3.predict(UnlabeledData(t));
    //predict most likely sentiment probabilities of t from model1, model2, and model3
    P1 = Model1.predict-prob(UnlabeledData(t));
    P2 = Model2.predict-prob(UnlabeledData(t));
    P3 = Model3.predict-prob(UnlabeledData(t));
    if( $S1 = S2 = S3$ ) & ( $P1 \geq \text{ThresholdValue}$  &  $P2 \geq \text{ThresholdValue}$  &  $P3 \geq \text{ThresholdValue}$ )then
      TrainingSet.Add(t);
      UnlabeledData.Remove(t);
    end
  end
end

```

3.3. Sentiment Classification

3.3.1. Dataset

We prepare our dataset to train a deep learning model and test the outcomes of the model using external benchmark datasets. Statistics about our corpora are shown in Table 6. We performed undersampling on the dataset by removing some of the tweets from the majority class randomly to match the number with the minority class. We considered 1 million tweets in each sentiment class to make the dataset balanced.

Table 6. Total tweets used to train the deep learning classifier from the last iteration.

	Positive	Negative	Neutral	Total
Training Set	1,013,576	1,013,576	1,013,576	3,040,728

We evaluated AraSenCorpus using external benchmark Arabic sentiment corpora and after each iteration. The benchmark datasets were SemEval 2017 and ASTD. The SemEval 2017 dataset contains tweets written in Arabic dialects and it is one of the most popular benchmarks for Arabic sentiment classification. ASTD is another dataset that contains 10,000 tweets written in modern standard Arabic and Egyptian dialect. Because we were performing two-way and three-way sentiment classification, we included the

tweets/reviews that were classified as positive, negative or neutral only. The statistics of these datasets are presented in Table 7.

Table 7. Statistics of benchmark datasets.

Dataset			Positive	Negative	Neutral	Total	Study
	Two	Balanced	743	743	–	1486	[3]
	Classes	Unbalanced	743	1142	–	1885	
SemEval 2017	Three	Balanced	743	743	743	2229	
	Classes	Unbalanced	743	1142	1444	3329	
ASTD	Two	Balanced	799	799	–	1598	[4]
	Classes	Unbalanced	799	1684	–	2483	
	Three	Balanced	799	799	799	2397	
	Classes	Unbalanced	799	1684	813	3296	

3.3.2. Experimental Setup

In this research, we performed two-way (positive and negative) and three-way (positive, negative, and neutral) sentiment classification. We trained a model using the LSTM deep learning classifier and tested the developed model on the benchmark datasets. The model was used with the hyper-parameter values as listed in Table 8.

Table 8. Hyperparameter of the deep learning classifiers.

Hyper Parameter	LSTM
Activation Function	softmax
Hidden Layers	6
Dropout rate	0.3
Learning rate	0.001
Number of epochs	10
Batch size	1024

3.3.3. Evaluation Metrics

All the results are reported using the F1-measure as follows:

$$F1 - \text{measure} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (2)$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (3)$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (4)$$

where the TP is the correctly predicted positive tweets, which means that the value of the actual class is positive and the value of the predicted class is also positive, FP is the actual class is negative and the predicted class is positive, and FN is the actual class is positive but the predicted class is negative. F1 is usually more useful because we have an uneven class distribution so it is better to look at both precision and recall measures. In our case, F1 score is used to measure the results and compare them with state-of-the-art systems.

4. Results

To evaluate the AraSenCorpus framework, we ran the framework using the manually annotated dataset (15,000 tweets) as the manually labeled input and extended it with the unlabeled dataset (more than 34 million tweets). In the last iteration, the full corpus contained more than 3.2 million and 4.5 million tweets for two-way and three-way sentiment classification, respectively. After obtaining each of the expanded annotated datasets, we

trained an LSTM deep learning model and tested it using the two benchmark datasets. The intuition behind these experiments was to show the quality of the proposed semi-automatic annotations approach.

The evaluation of the AraSenCorpus corpus was carried out using two benchmark datasets to evaluate the effectiveness of the semi-supervision annotation. For this, we used SemEval 2017 and ASTD datasets. All experiments on these datasets were performed using two-way classification (positive and negative classes) and three-way sentiment classification (positive, negative, and neutral) using both balanced and unbalanced datasets.

Figure 5 shows the F1-score results for two-way sentiment classification over two benchmark datasets after performing the expansion of the manual labeled dataset. It is clearly depicted that the increasing of labeled tweets with high confidence leads to better results. For example, the classification starts with the manually annotated dataset (15,000 tweets). It gives a low score in comparison with the results obtained in the third iteration (3.2 million tweets) using both balanced and unbalanced datasets. The last iteration shows significant results overall for the benchmark datasets.

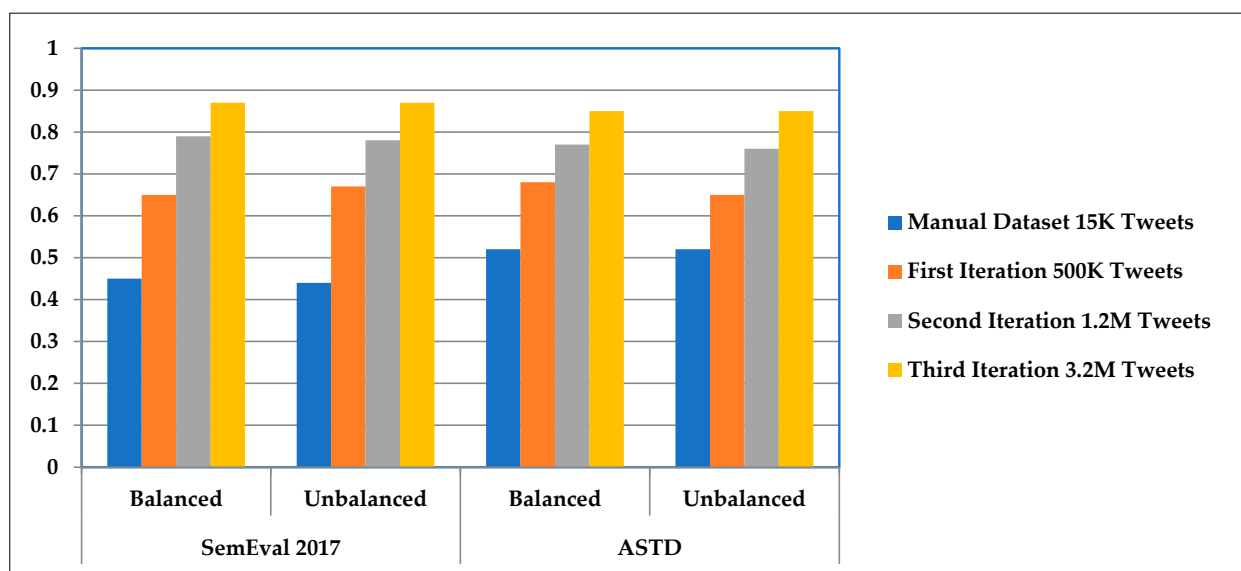


Figure 5. Two-way classification results (F1-Score): the graph shows the improvement of results after each expansion process.

Figure 6 shows the F1-score results for three-way sentiment classification. The obtained results in the third iteration are higher than the results obtained in the manual, first iteration, and second iteration datasets. It is clearly shown that the results obtained from the two-way sentiment classification are higher than the results obtained from the three-way sentiment classification. This is due to the existence of the neutral class in the three-way sentiment classification.

The above results show the performance of our system on the benchmark datasets. Next, we compare the best-obtained results by our system with previous studies in both two-way and three-way classification, as shown in Tables 9 and 10, respectively. For the two-way classification, our system outperforms two recent studies in all datasets, as shown in Table 9. Our system improves the sentiment classification results from 80.37% to 87.4% using the SemEval 2017 dataset and from 79.77% to 85.2% using the ASTD dataset.

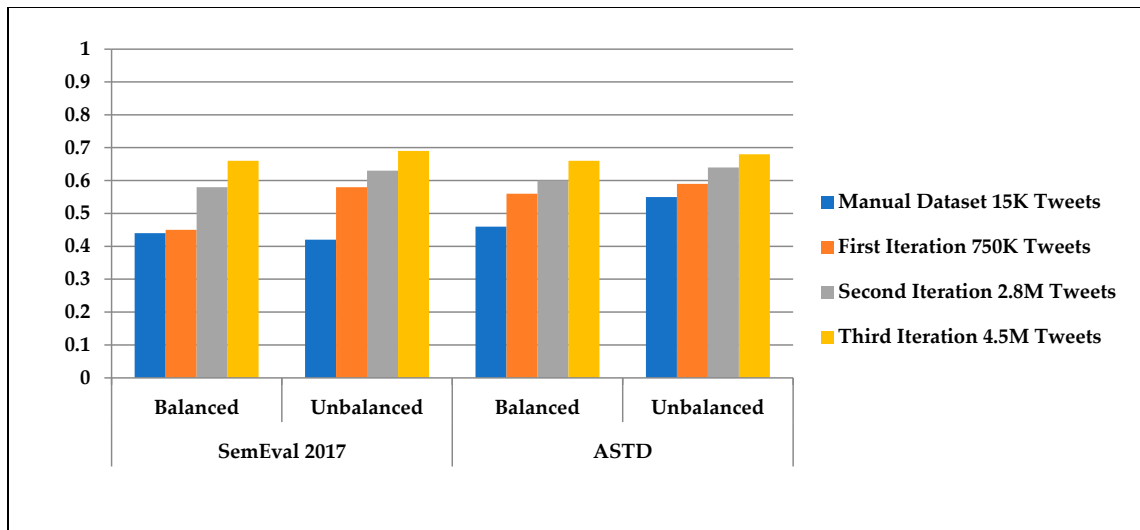


Figure 6. Three-way classification results (F1-Score): the graph shows the improvement of results after each expansion process.

Table 9. Two-way classification comparison of existing methods with our models on the same datasets (F1-Score).

Dataset	Al-Twairsh, N.; Al-Negheimish, H.2019 [20]	Farha, Ibrahim et al., 2019 [41]	Our Models
SemEval 2017	80.37	63	87.4
ASTD	79.77	72	85.2

The bold values represent the best obtained results.

Table 10. Three-way sentiment classification comparison of existing methods with our models (F1-Score).

Dataset	Farha, Ibrahim et al., 2019 [41]	Our Models
SemEval 2017	63.38	69.4
ASTD	64.10	68.1

The bold values represent the best obtained results.

In a three-way classification, our system also achieves the best results, as shown in Table 10. Our system gives 69.4% accuracy for the SemEval 2017 dataset while the best system gives 63.38% using F1-score. It also achieved 68.1% accuracy with the ASTD dataset, while the best performing system achieved 64.10%.

5. Discussion and Future Work

The major finding in our research is a large-scale Arabic sentiment corpus named “AraSenCorpus”. The introduced techniques in building our corpus help in reducing human effort and automate the annotation process. The corpus contains 4.5 million tweets and achieves potential results when compared with state-of-the-art system using the same benchmark datasets.

The sentiment classification on two-way classification using our corpus improves the results by 7% and 5% using the SemEval 2017 and ASTD benchmark datasets, respectively. The three-way classification using our corpus also improves the classification results by 6% and 4% using both benchmark datasets, respectively. The tweets in our corpus are mostly subjective; as our corpus was collected using terms/phrases from Arabic sentiment lexicons. From the results and analysis, we observe that the two studies that we compared our results with used a mixed of subjective and objective tweets. Along with this reason,

our collected tweets spanned 14 years to ensure the coverage of different topics, while in the two studies used in the comparison, the tweets spanned 2 months and 3 years, respectively.

This semi-supervised learning approach proves that the expansion of a few samples that have been carefully annotated will lead to better results, as shown in the presented corpus. As the expanded corpus has millions of tweets, we used LSTM deep learning classifier which gives better results with sufficient data, as in our corpus.

AraSenCorpus is limited by the weaknesses of self-learning, such as skewed class distributions and error propagation. To overcome this problem, we selected tweets from the unlabeled dataset above or greater than a certain threshold, to be added to the training set. Other alternative semi-supervised approaches, such as co-training, can be used.

Improvements to AraSenCorpus could include improvement of the guidelines of the manual annotation, which could improve the classification results. Further preprocessing steps will be added to improve the overall performance of the sentiment classification.

6. Conclusions

In this paper, we present AraSenCorpus, a semi-supervised framework to annotate a large corpus for Arabic sentiment analysis. In doing so, we use the self-learning approach by training three neural network models to expand the training set with new data after applying a certain confidence threshold. We use a manually annotated dataset containing 15,000 tweets to annotate more than 4 million tweets. The sentiment classification results obtained using the LSTM deep learning classifier from AraSenCorpus outperforms the existing state-of-the-art model. The text of the corpus is a mix of modern standard Arabic and some of the Arabic dialects including Gulf, Yemeni, Egyptian, Iraqi, and Levantine dialects.

The proposed sentiment corpora in the literature are either small in size, not publicly available, or were annotated using fully automatic annotation methods. We overcome such issues by offering a large-scale sentiment corpus that is freely available for research purposes. The semi-supervised annotation method helps in automating the annotation and reducing the human effort during the annotation process.

As there is a lack of large sentiment corpora in the Arabic language, our corpus contributes in tackling the scarcity of Arabic corpora for sentiment analysis. The developed corpus contains more than 4.5 million tweets that were labeled into three classes (positive, negative, and neutral).

In the future, we plan to add more dialects, such as Sudanese, Moroccan, Algerian, and Tunisian dialects by training models on freely available sentiment corpora in these dialects. We will also experiment with more classification algorithms and use different inputs such as BERT to improve the sentiment classification of Arabic. Along with this, we will include more statistical analysis on the obtained and future results. Our intention is to build the largest sentiment corpus that covers modern standard Arabic and all dialectical Arabic.

Author Contributions: Conceptualization, A.A.-L.; Formal analysis, A.A.-L.; Funding acquisition, H.F.A.; Methodology, A.A.-L.; Project administration, M.S.; Resources, A.A.-L.; Software, A.A.-L.; Supervision, M.S.; Visualization, A.A.-L.; Writing—original draft, A.A.-L.; Writing—review & editing, M.S., H.F.A. and A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Prince Sultan University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code datasets are freely available for research purposes in this link: <https://github.com/yemen2016/AraSenCorpus> (accessed on 26 February 2021).

Acknowledgments: Authors are thankful to Prince Sultan University, Saudi Arabia for providing the fund to carry out the work.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Hovy, E.; Lavid, J. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *Int. J. Transl.* **2010**, *22*, 13–36.
2. Horbach, A.; Thater, S.; Steffen, D.; Fischer, P.M.; Witt, A.; Pinkal, M. Internet corpora: A challenge for linguistic processing. *Datenbank-Spektrum* **2015**, *15*, 41–47. [[CrossRef](#)]
3. Rosenthal, S.; Farra, N.; Nakov, P. SemEval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017.
4. Nabil, M.; Aly, M.; Atiya, A. Astd: Arabic sentiment tweets dataset. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015.
5. fastText. Available online: <https://fasttext.cc/> (accessed on 2 March 2021).
6. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of tricks for efficient text classification. *arXiv* **2016**, arXiv:1607.01759.
7. Rao, A.; Spasojevic, N. Actionable and political text classification using word embeddings and lstm. *arXiv* **2016**, arXiv:1607.02501.
8. Baly, R.; El-Khoury, G.; Moukalled, R.; Aoun, R.; Hajj, H.; Shaban, K.B.; El-Hajj, W. Comparative evaluation of sentiment analysis methods across Arabic dialects. *Procedia Comput. Sci.* **2017**, *117*, 266–273. [[CrossRef](#)]
9. Al-Laith, A.; Shahbaz, M. Tracking sentiment towards news entities from arabic news on social media. *Future Gener. Comput. Syst.* **2021**, *118*, 467–484. [[CrossRef](#)]
10. Aly, M.; Atiya, A. Labr: A large scale arabic book reviews dataset. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, 4–9 August 2013.
11. ElSahar, H.; El-Beltagy, S.R. Building large arabic multi-domain resources for sentiment analysis. In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt, 14–20 April 2015; Springer: Berlin/Heidelberg, Germany, 2015.
12. Elnagar, A.; Einea, O. Brad 1.0: Book reviews in arabic dataset. In Proceedings of the 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), Agadir, Morocco, 29 November–2 December 2016; IEEE: Piscataway, NJ, USA, 2016.
13. Elnagar, A.; Lulu, L.; Einea, O. An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis. *Procedia Comput. Sci.* **2018**, *142*, 182–189.
14. Elnagar, A.; Khalifa, Y.S.; Einea, A. Hotel Arabic-reviews dataset construction for sentiment analysis applications. In *Intelligent Natural Language Processing: Trends and Applications*; Springer: Cham, Switzerland, 2018; pp. 35–52.
15. Guellil, I.; Adeel, A.; Azouaou, F.; Hussain, A. Sentialg: Automated corpus annotation for algerian sentiment analysis. In Proceedings of the International Conference on Brain Inspired Cognitive Systems, Xi'an, China, 7–8 July 2018; Springer: Berlin/Heidelberg, Germany, 2018.
16. Gamal, D.; Alfonse, M.; El-Horbaty, E.S.M.; Salem, A.B.M. Twitter benchmark dataset for Arabic sentiment analysis. *Int. J. Mod. Educ. Comput. Sci.* **2019**, *11*, 33. [[CrossRef](#)]
17. Abdellaoui, H.; Zrigui, M. Using tweets and emojis to build tead: An Arabic dataset for sentiment analysis. *Comput. Sist.* **2018**, *22*, 777–786. [[CrossRef](#)]
18. Dahou, A.; Xiong, S.; Zhou, J.; Haddoud, M.H.; Duan, P. Word embeddings and convolutional neural network for arabic sentiment classification. In Proceedings of the Coling 2016, the 26th International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016.
19. Abo, M.E.M.; Shah, N.A.K.; Balakrishnan, V.; Kamal, M.; Abdelaziz, A.; Haruna, K. SSA-SDA: Subjectivity and sentiment analysis of sudanese dialect Arabic. In Proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS), Aljouf, Saudi Arabia, 10–11 April 2019; IEEE: Piscataway, NJ, USA, 2019.
20. Al-Twairesh, N.; Al-Negheimish, H. Surface and deep features ensemble for sentiment analysis of arabic tweets. *IEEE Access* **2019**, *7*, 84122–84131. [[CrossRef](#)]
21. Al-Twairesh, N.; Al-Khalifa, H.; Al-Salman, A.; Al-Ohali, Y. Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. *Procedia Comput. Sci.* **2017**, *117*, 63–72. [[CrossRef](#)]
22. Alqarafi, A.; Adeel, A.; Hawalah, A.; Swingler, K.; Hussain, A. A Semi-supervised Corpus Annotation for Saudi Sentiment Analysis Using Twitter. In Proceedings of the International Conference on Brain Inspired Cognitive Systems, Xi'an, China, 7–8 July 2018; Springer: Berlin/Heidelberg, Germany, 2018.
23. Brum, H.B.; Nunes, M.D.G.V. Semi-supervised Sentiment Annotation of Large Corpora. In Proceedings of the International Conference on Computational Processing of the Portuguese Language, Canela, Brazil, 24–26 September 2018; Springer: Berlin/Heidelberg, Germany, 2018.
24. Iosifidis, V.; Ntoutsi, E. Large scale sentiment learning with limited labels. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017.
25. Amazon Mechanical Turk. Available online: <https://www.mturk.com> (accessed on 2 March 2021).
26. Alahmary, R.M.; Al-Dossari, H.Z.; Emam, A.Z. Sentiment analysis of Saudi dialect using deep learning techniques. In Proceedings of the 2019 International Conference on Electronics, Information, and Communication (ICEIC), Auckland, New Zealand, 22–25 January 2019; IEEE: Piscataway, NJ, USA, 2019.
27. Baly, R.; Khaddaj, A.; Hajj, H.; El-Hajj, W.; Shaban, K.B. Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets. *arXiv* **2019**, arXiv:1906.01830, 2019.

28. CrowdFlowerplatform. Available online: <https://appen.com/> (accessed on 2 March 2021).
29. Rahab, H.; Zitouni, A.; Djoudi, M. SANA: Sentiment analysis on newspapers comments in Algeria. *J. King Saud Univ. Comput. Inf. Sci.* **2019**. [[CrossRef](#)]
30. Al-Thubaity, A.; Alharbi, M.; Alqahtani, S.; Aljandal, A. A Saudi dialect twitter corpus for sentiment and emotion analysis. In Proceedings of the 2018 21st Saudi Computer Society National Computer Conference (NCC), Riyadh, Saudi Arabia, 25–26 April 2018; IEEE: Piscataway, NJ, USA, 2018.
31. Atoum, J.O.; Nouman, M. Sentiment analysis of Arabic Jordanian dialect tweets. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 256–262. [[CrossRef](#)]
32. Al-Moslemi, T.; Albared, M.; Al-Shabi, A.; Omar, N.; Abdullah, S. Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis. *J. Inf. Sci.* **2018**, *44*, 345–362. [[CrossRef](#)]
33. Oussous, A.; Benjelloun, F.Z.; Lahcen, A.A.; Belfkih, S. ASA: A framework for Arabic sentiment analysis. *J. Inf. Sci.* **2020**, *46*, 544–559. [[CrossRef](#)]
34. Mdhaffar, S.; Bougares, F.; Esteve, Y.; Hadrich-Belguith, L. Sentiment analysis of Tunisian dialects: Linguistic resources and experiments. In Proceedings of the Third Arabic Natural Language Processing Workshop (WANLP 2017), Valencia, Spain, 3–4 April 2017.
35. Abdul-Mageed, M.; Diab, M.T. AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In Proceedings of the LREC 2012, Istanbul, Turkey, 21–27 May 2012.
36. Mourad, A.; 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, GA, USA, 13–14 June 2013.
37. Abdulla, N.A.; Ahmed, N.A.; Shehab, M.A.; Al-Ayyoub, M. Arabic sentiment analysis: Lexicon-based and corpus-based. In Proceedings of the 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan, 3–5 December 2013; IEEE: Piscataway, NJ, USA, 2013.
38. Catal, C.; Nangir, M. A sentiment classification model based on multiple classifiers. *Appl. Soft Comput.* **2017**, *50*, 135–141. [[CrossRef](#)]
39. Alharbi, F.R.; Khan, M.B. Identifying comparative opinions in Arabic text in social media using machine learning techniques. *SN Appl. Sci.* **2019**, *1*, 213. [[CrossRef](#)]
40. Al-Laith, A.; Alenezi, M. Monitoring People’s Emotions and Symptoms from Arabic Tweets during the COVID-19 Pandemic. *Information* **2021**, *12*, 86. [[CrossRef](#)]
41. Farha, I.A.; Magdy, W. Mazajak: An online Arabic sentiment analyser. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 28 July–2 August 2019.
42. Mulki, H.; Haddad, H.; Gridach, M.; Babaoglu, I. Empirical evaluation of leveraging named entities for Arabic sentiment analysis. *arXiv* **2019**, arXiv:1904.10195. [[CrossRef](#)]
43. El-Beltagy, S.R. NileULex: A phrase and word level sentiment lexicon for Egyptian and modern standard Arabic. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), Portorož, Slovenia, 23–28 May 2016.
44. Abdulla, N.A.; Ahmed, N.A.; Shehab, M.A.; Al-Ayyoub, M.; Al-Kabi, M.N.; Al-rifai, S. Towards improving the lexicon-based approach for Arabic sentiment analysis. *Int. J. Inf. Technol. Web Eng. (IJITWE)* **2014**, *9*, 55–71. [[CrossRef](#)]
45. Number of Monthly Active Twitter Users Worldwide from 1st Quarter 2010 to 1st Quarter 2019. Available online: <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/> (accessed on 2 March 2021).
46. Mohammad Salameh, S.M.M.; Kiritchenko, S. Arabic Sentiment Analysis and Cross-lingual Sentiment Resources. Available online: <https://saifmohammad.com/WebPages/ArabicSA.html> (accessed on 2 March 2021).
47. Elsahar, H. Large Multi-Domain Resources for Arabic Sentiment Analysis. Available online: <https://github.com/hadyelsahar/large-arabic-sentiment-analysis-resources> (accessed on 2 March 2021).
48. NileULex. Available online: <https://github.com/NileTMRG/NileULex> (accessed on 2 March 2021).
49. MASC. Available online: <https://github.com/almosmi/masc> (accessed on 2 March 2021).
50. Salameh, M.; Mohammad, S.; Kiritchenko, S. Sentiment after translation: A case-study on Arabic social media posts. In Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015.
51. Youssef, M.; El-Beltagy, S.R. MoArLex: An Arabic sentiment lexicon built through automatic lexicon expansion. *Procedia Comput. Sci.* **2018**, *142*, 94–103. [[CrossRef](#)]
52. Torre, I.G.; Luque, B.; Lacasa, L.; Kello, C.T.; Hernández-Fernández, A. On the physical origin of linguistic laws and lognormality in speech. *R. Soc. Open Sci.* **2019**, *6*, 191023. [[CrossRef](#)] [[PubMed](#)]
53. Sicilia-Garcia, J.; Ming, E.I.; Smith, F.J. Extension of Zipf’s law to words and phrases. In Proceedings of the COLING 2002: The 19th International Conference on Computational Linguistics, Taipei, Taiwan, 26–30 August 2002.
54. Fralick, S. Learning to recognize patterns without a teacher. *IEEE Trans. Inf. Theory* **1967**, *13*, 57–64. [[CrossRef](#)]