Arabic Social Media Sentiment Analysis Using RCNN

Workshop

Rania Abd EL Monam Mohamed

Department of Computer Sciences
Faculty of Graduate Studies for Statistical Research
Cairo University, Egypt

11/12/2019

Agenda

1. Introduction.
2. Proposed Corpus.
3. Embedding Layer.
4. Architecture of RCNN Model.
5. Publications.
6. RCNN Code.

Agenda

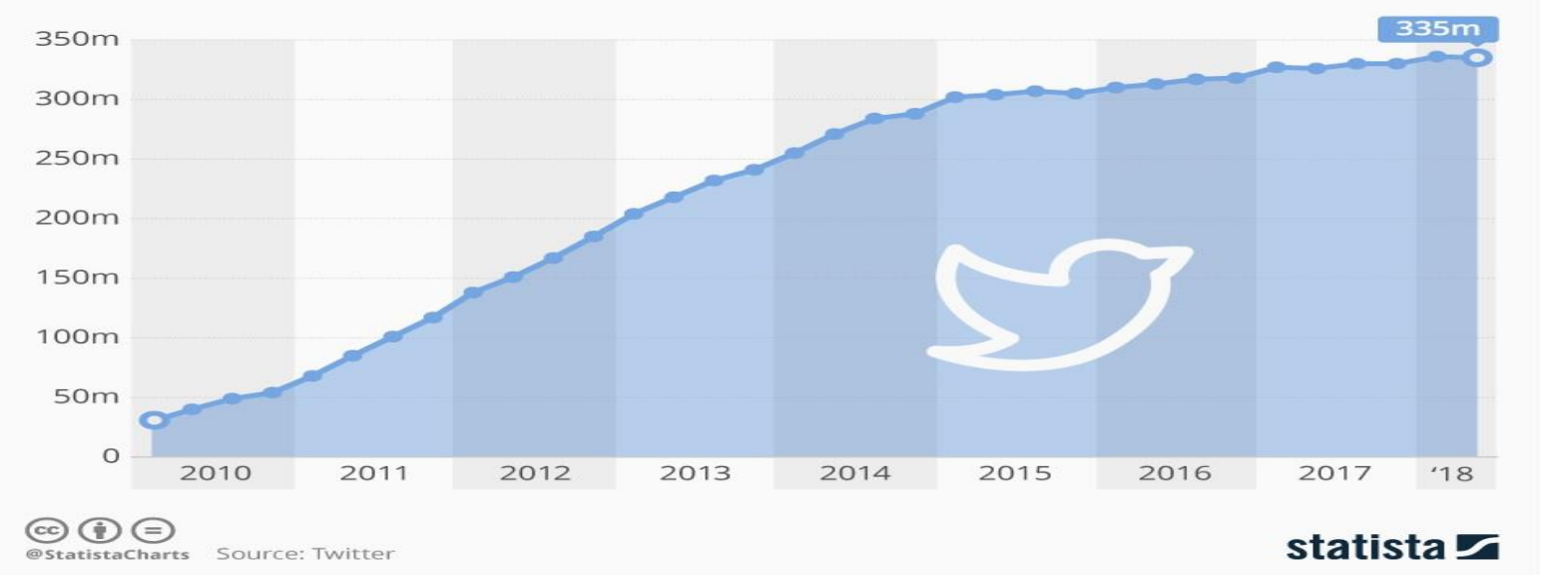
1. Introduction.
2. Proposed Corpus.
3. Embedding Layer.
4. Architecture of RCNN Model.
5. Publications.
6. RCNN Code.

Social Media

- Social Media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc.
- Social Media are considered an excellent source of information and can provide opinions, thoughts, and insights toward various important topics.
- It is expected that the estimated amount of data on the web will be about 40 trillion gigabytes, in 2020 [1].

1- Gantz, J. and D. Reinsel, *Digital Universe Study.* Extracting Value from Chaos, IDC Go-to-Market Services, 2011.

- Twitter is one of the most popular, widely used social media.
- Official Twitter statistics show that 335 million monthly users actively used Twitter in 2018.



Number of monthly active twitter users around the world from 2010 to 2018 [2]

2- Twitter, *https://www.businessofapps.com/data/twitter-statistics/*. 2018.

Sentiment Analysis

- Sentiment Analysis (SA) also known as review mining.
- SA can be defined as a process that automates mining of attitudes, opinions and emotions from text.
- SA is the process of determining whether a piece of writing is positive ☺ or negative ☹.
- SA refers to the process of deriving high-quality information from text.

- SA plays a significant role in our daily decision making process.
- SA helps in achieving various goals like observing public mood regarding political topics, market intelligence, new product sales prediction, the measurement of customer satisfaction and many more.
- SA can be divided into three levels, namely

✓ Document Level

✓ Sentence Level.

✓ Aspect Level (also known as word- or feature- level).

Arabic Contents

- Arabic language is one of six official languages of the United Nations.
- Arabic is the official language of 21 countries, and it's the major language in several areas of the world.
- Arabic language is classified into three types;

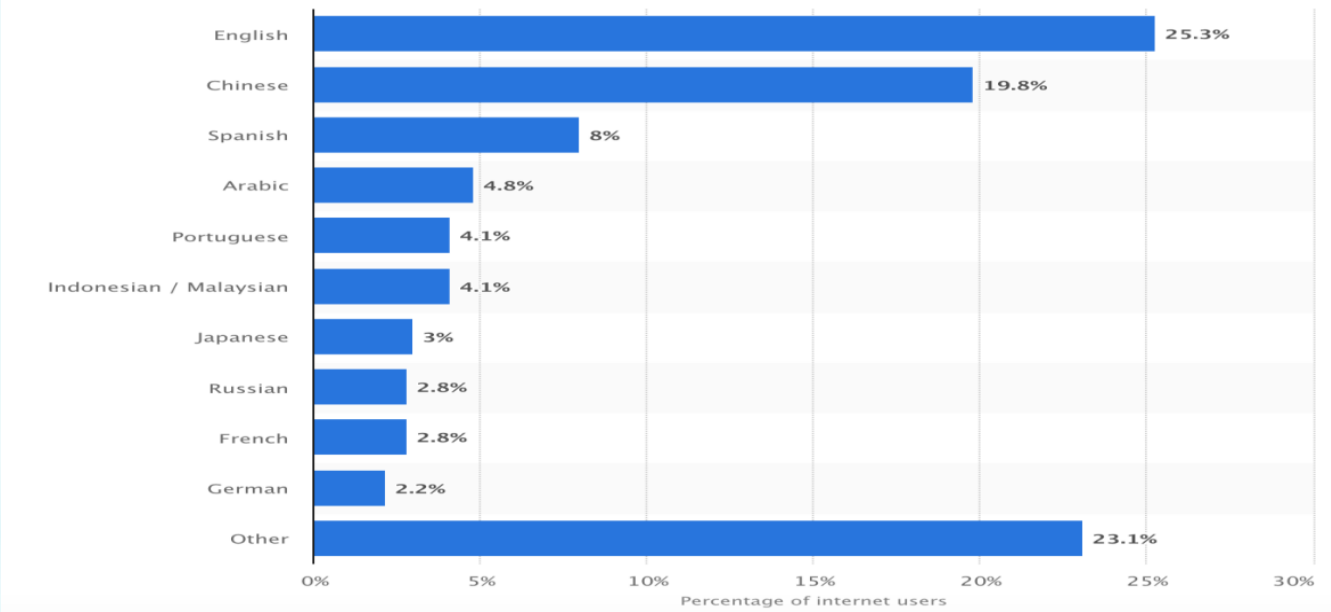
✓ Classical Arabic ("AL-Qur'an")

✓ Modern Standard Arabic (formal communications, television, radio, news, education)

✓ Colloquial Arabic (shopping, chatting or in their homes.)

Arabic Contents

- Arabic is currently ranked as the forth language used in the web, and there are about 168 million of Arabic Internet user [5].



Top 10 Languages used on the internet today

5- Languages, https://www.internetworldstats.com/stats7.htm , 2018.

Colloquial Arabic language

Analyzing Arabic contents is a challenging task due to:

- Unstructured language (informal language).
- Spelling inconsistencies.
- Slang words. Such as (كدا, ولو , كويس, يلا)
- Idiomatic expressions.

  Such as (الطيور على اشكالها تقع, التكرار يعلم الحمار, بعد ما شاب ودوه الكتّاب,
  هاك الشبل من ذاك الأسد, شحات ونزهي, طباخ السم بيدوقه)
- Colloquial expressions.

Such as (ناس مابتجيش الا بالعين الحمرا, مفيش فايدة, في المشمش, يا خبر أبيض,
يا سلام, زي السمن على العسل, يسرق الكحل من العين, زي القطط تاكل وتنكر)

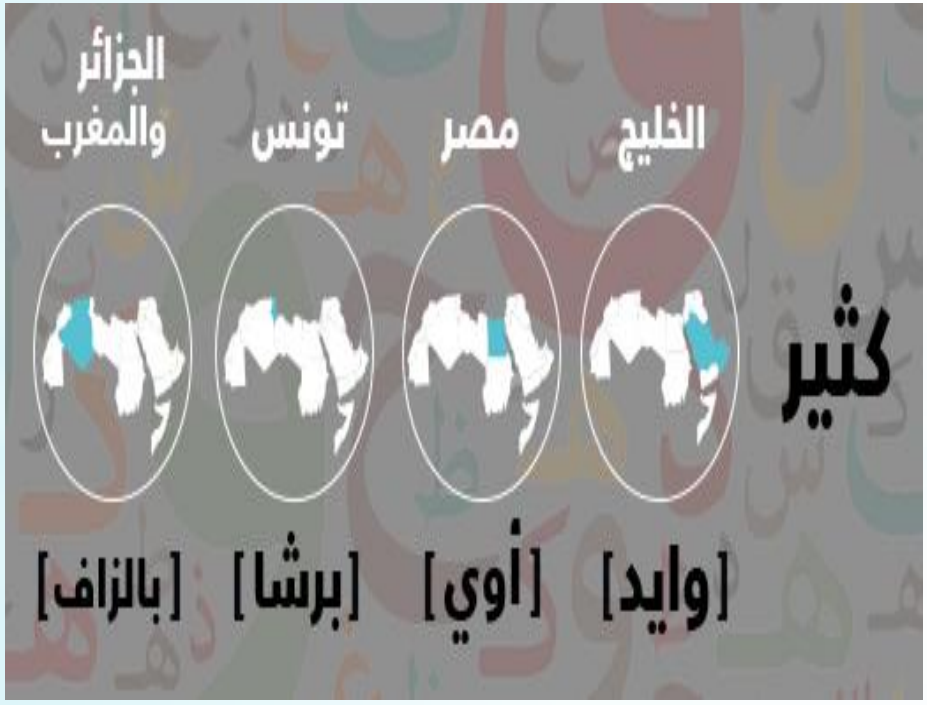
- Emoticons. Such as (🙂, 😀 , 😢 , 😅 , 😂)
- The tendency to repeat letters in writing to convey feelings.

Such as(كتيررررررررر)

Colloquial Arabic language

- **"Colloquial Arabic" (CA) is most widely spoken and written in daily live conversations with different dialects.**

(colloquial Arabic varies from region to another region, from country to other countries, even from even state to another.)

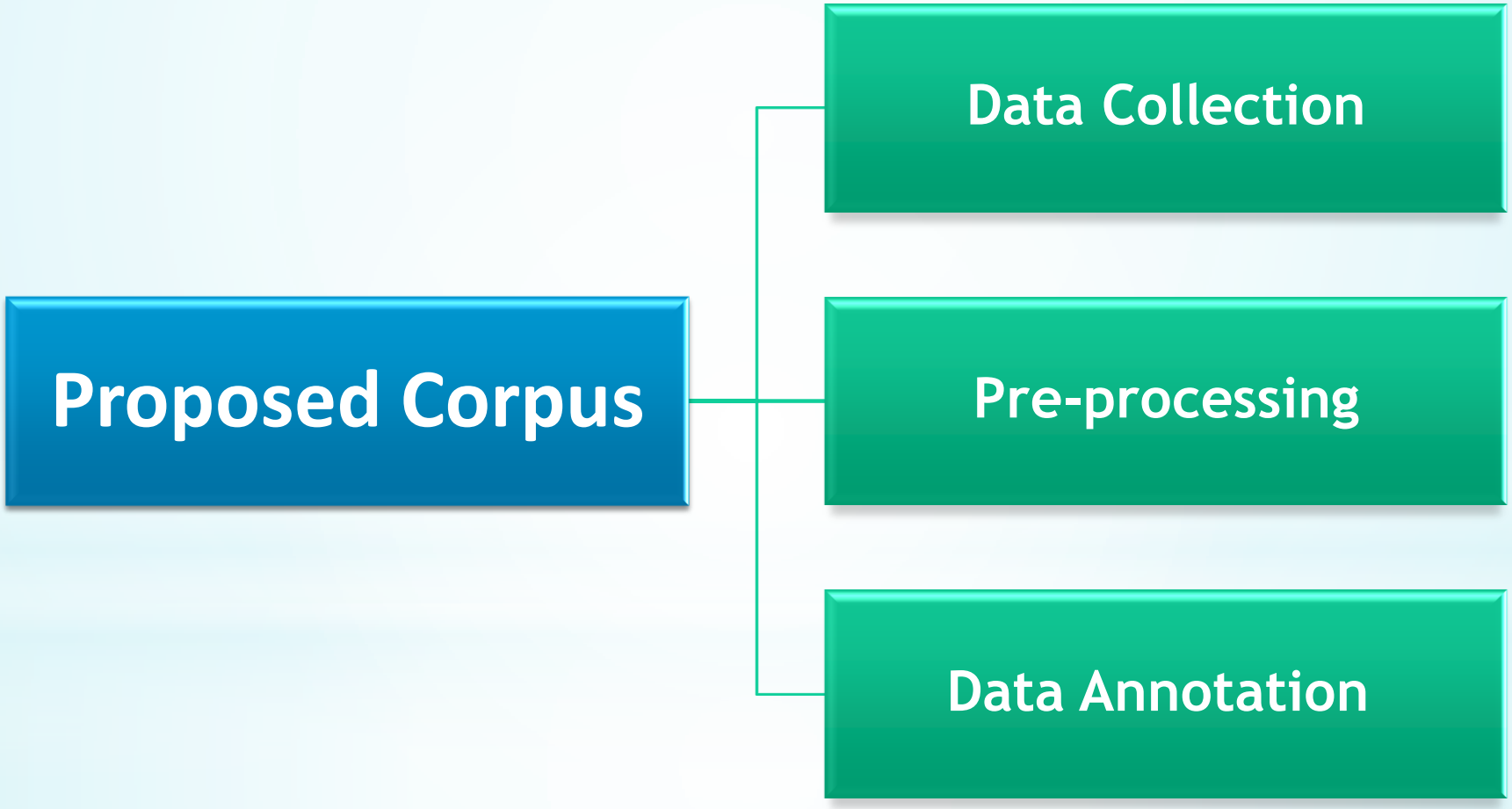


Colloquial dialect	Sentence
Jordanian	أيش بدك
Egyptian	عايز ايه
Saudi	وش تبغى
Tunisian	شنو تحب
Algerian	واش تحب

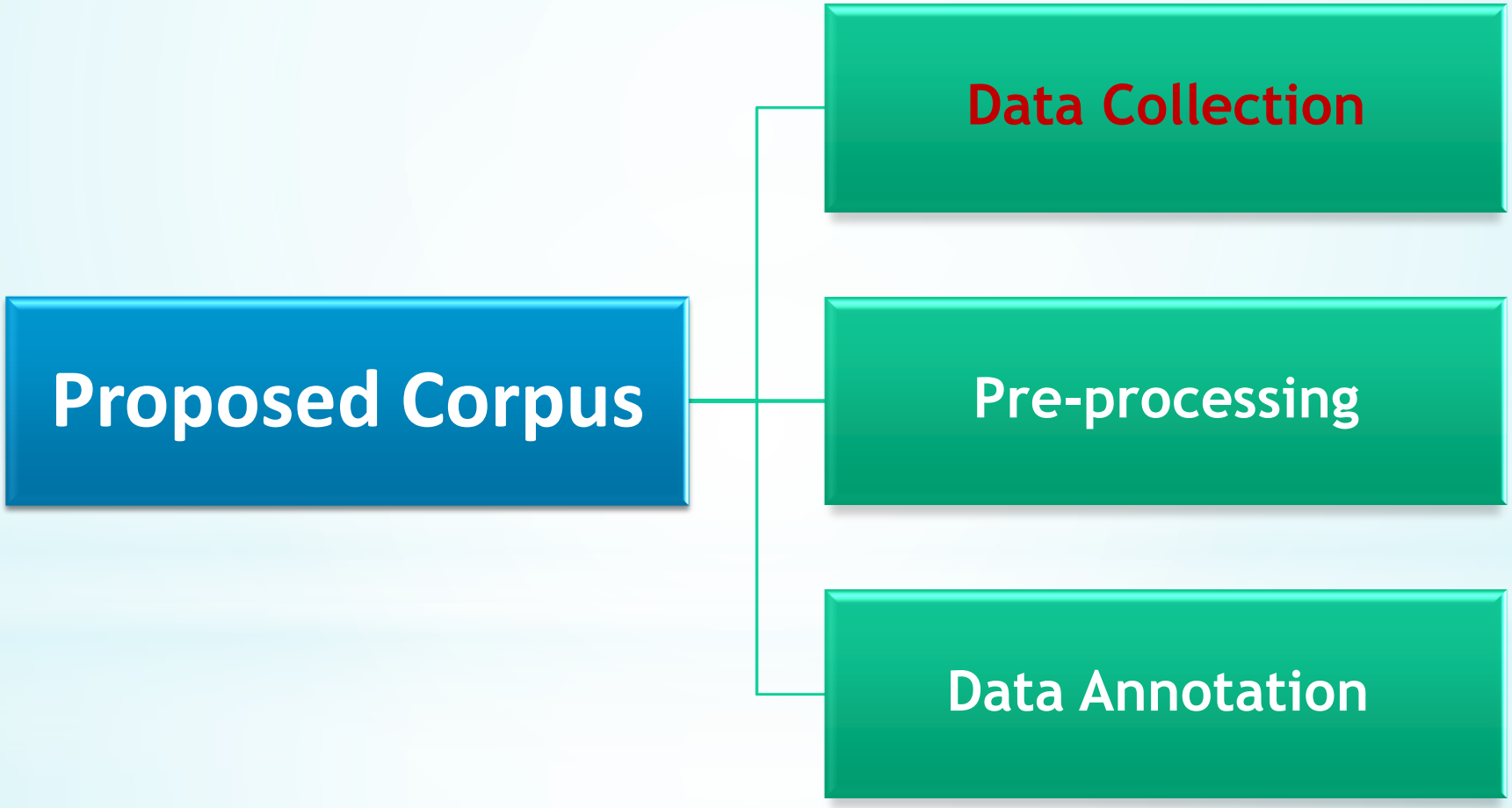
Agenda

1. Introduction.
2. Proposed Corpus.
3. Embedding Layer.
4. Architecture of RCNN Model.
5. Publications.
6. RCNN Code.

Proposed Corpus (Egyptian Corpus)



Proposed Corpus (Egyptian Corpus)



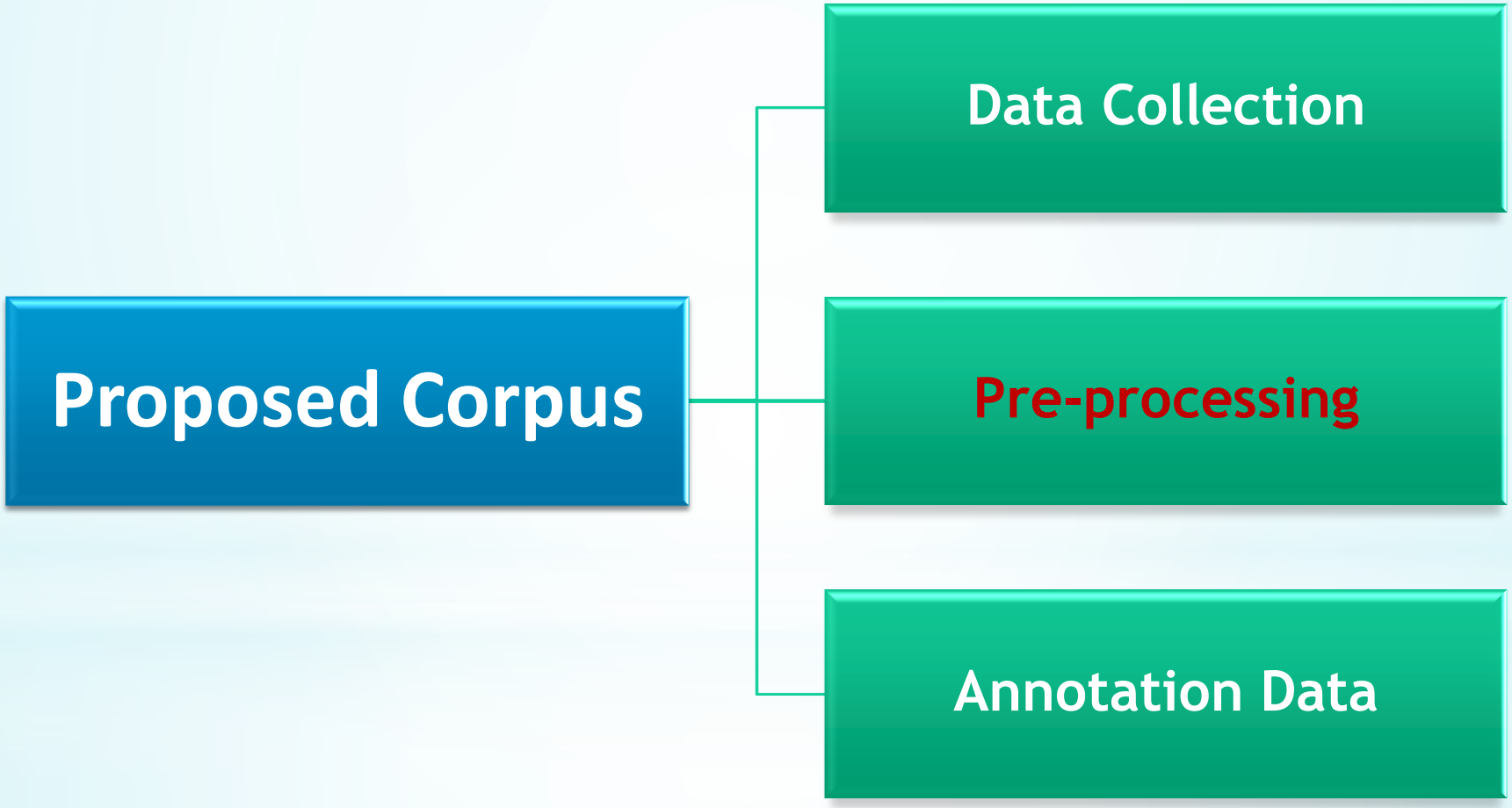
Proposed Corpus (Egyptian Corpus)

- ✓ Tweets was collected using twitter APIs.
- ✓ The collected tweets span the period from April 11, 2015 to December 12, 2015. In this initial stage, a collection of 2,408,128 noisy raw tweets were gathered.

37540253683712	@AppleIgenius	NULL19 Dec 12	عزيزي الفرج بيقولو الصبر مفتاحك .. انت غيرت الكالون ولا اية يابرنس :D
37539339321344	@AAlAnezi	NULL19 Dec 12	@farisf9 غلبت سيبويه بالكناية والسجع يا بو نورة :)
37533094027264	@ibrahimaljameel	NULL19 Dec 12	جلسات وناسة ٢٠١٠ مصخرة :) وشكراً #بس_أقول
37526437642241	@ghamd07sss	NULL19 Dec 12	....(^_*).... ليس كل من تعلم الاحرف :) اتقن الكلامالجميع يكتب والقليل يقرا :(والنادر يفهم
37523493261312	@AyaAliKhalid	NULL19 Dec 12	:D لما تبقى راكب العربيه وأغنيه حزينه تشتغل بتفضل تبص من الشباك وإنت زعلان كإنك بتمثل الفيلم
37522495012866	@saad1019	NULL19 Dec 12	انقفدَ خسسأارهمْ فَ الاسنغناَء عن البعض :) * حيأهه جديده ♡ ------------
37518422351872	@NojoudQ	NULL19 Dec 12	#نصائح_للمذاكرة_قبل_الاختبار اتعادي جداً تجمع مخدات قدامك وتشرح لها :))) أهم شي النتيجه في الورقة
37511304593408	@ANTi7sceularism	NULL19 Dec 12	ما حدث لاحمد عرفه اليوم هو رسالة وجث نبض لكل ادمنات الصفحات الاسلامية .نحب نقول للداخلية طز فيكم :)مكملين
37511254274049	@dee50_	NULL19 Dec 12	! ♥:) ليتَ رَائحةُ المَطَر ، تُباع فِي زُجاجَة عَطر
37507210960896	@muhamedrefaay2	NULL19 Dec 12	مصر كلها عايزه فولورز :-) #كلنا_مصريين_وبنعمل_فولو_لبعض
37489708146688	@m_uaz	NULL19 Dec 12	:) في المستشفىَ على اليمين رجلٌ يستخرج شهأدة ميلأد إبنه و على يسأره آخر يستخرج شهأدة وفأةِ أبيه ! .. مشهَد يختصِر الحيأهه
37485157310464	@budrees1985	NULL19 Dec 12	@alaoy_alaoyrf اصدقائياصدرتاني تطبيق ايفون مجاناً الرجاء تحميله وتقييمه itunes.apple.com
37483857084417	@adel_200895	NULL19 Dec 12	الجيزه ، البحيره ، المنيا ، القليوبيه و أخيرا المنوفيه كلهم مهمين :-)أنا بس كنت بسلط الضوء على المنوفيه في... fb.me/2c4
37482565246978	@Ghada_Nasser_	NULL19 Dec 12	طب الظهر كام ركعة :P ?? – لا اله الا الله [[asa7beee]] ask.fm/a/1kafco44
37474134671361	@tscor2	NULL19 Dec 12	ذاك الشبل من ذاك الأسد .. :) انها الوراثة يا عزيزي...AB fb.me/zahAZxto
37473748803584	@iewsh	NULL19 Dec 12	.:) في آخر الليل ؛ إن لم تكن مصلياً كُن مستغفراً
37472477929472	@DalyaFarrag	NULL19 Dec 12	:D !.. وراء كل امرأة مجنونة .. رجل قاعد يغيظ فيها
37472217874433	@Tahani_nasser	NULL19 Dec 12	:p #ماذا_تقول_لمن_يراقبك_على_تويتر دققوا منيح لا يفوتكم شيء
37462524866560	@fahadalenazi5	NULL19 Dec 12	:) يا زين ليلة الخميس
37461128159232	@th20133	NULL19 Dec 12	':) إن تحبني انت مخير لأمجبر ولكن إن تحترمني انت مجبر ؤلا مخير
37455486799872	@Mahakth	NULL19 Dec 12	! :) انا عندي عقده من الأصوات العاليه أو اللي يصارخ ، ينرفزني
37452643065856	@_aaa112	NULL19 Dec 12	:) البكاء بلا اسباب !*أفاقه من سُبات طويل ، ل كلام مُتكدس
37449996472320	@tharwat_hassan	NULL19 Dec 12	ما هي أكثر صفة تعجبك من صفات المرأة؟ – el3'baa2 :D :P :P :P :P ask.fm/a/118:
37445789581313	@ArwaTarek33	NULL19 Dec 12	:D المجد لكل واحد مايعرفش عنده ايه فى المنهج لحد دلوقتى

Original tweets

Proposed Corpus (Egyptian Corpus)

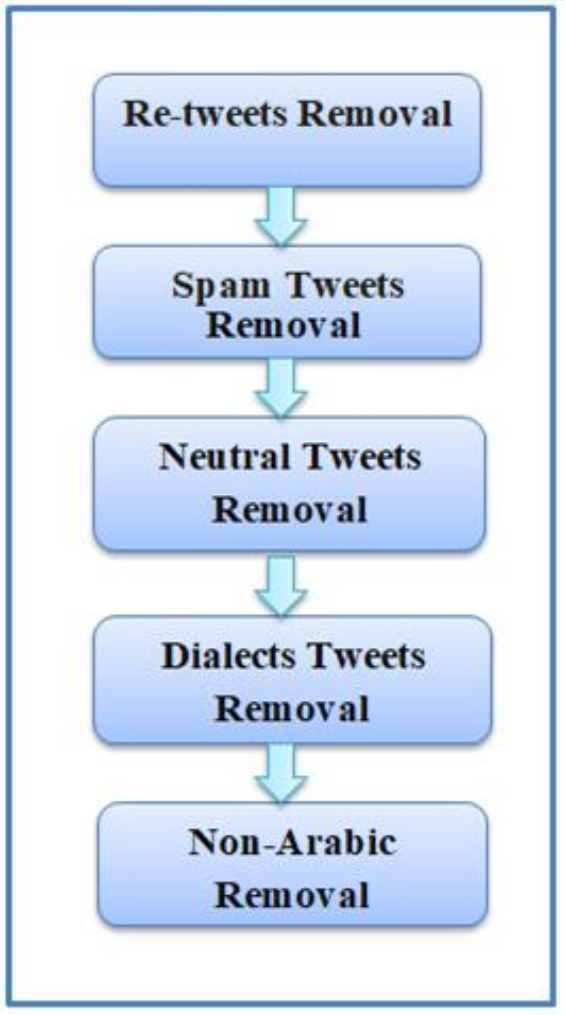


Preprocessing (Egyptian Corpus)

- **Phase 1:** Filtering Data
- **Phase 2:** Cleaning Data
- **Phase 3:** Normalization Data

Phase 1: Filtering Data

- After collecting the data from Twitter, we found that the collected tweets have many problems that will affect the annotation process accuracy.
- This phase consisted of five steps manually:



Tweet filtering phase

Phase 2: Cleaning Data

- ✓ This phase is important since it deals with the noisy nature of tweets.
- ✓ data cleaning deals manually:

  Such as (hashtag (#), username (@username),emoticons (=D, XD), time stamps and URL)

- ✓ Some tweets may contain:

  (missing spelling words, missing letters and written incorrectly)

- ✓ This problem is handled by correcting these words and wrong characters.
- ✓ Some words are combined together due to space limitation of Twitter. This problem is handled by adding spaces between words.

Phase 3: Normalization Data

- Normalization is the process of transforming the Arabic text in order to be consistent.

For example:

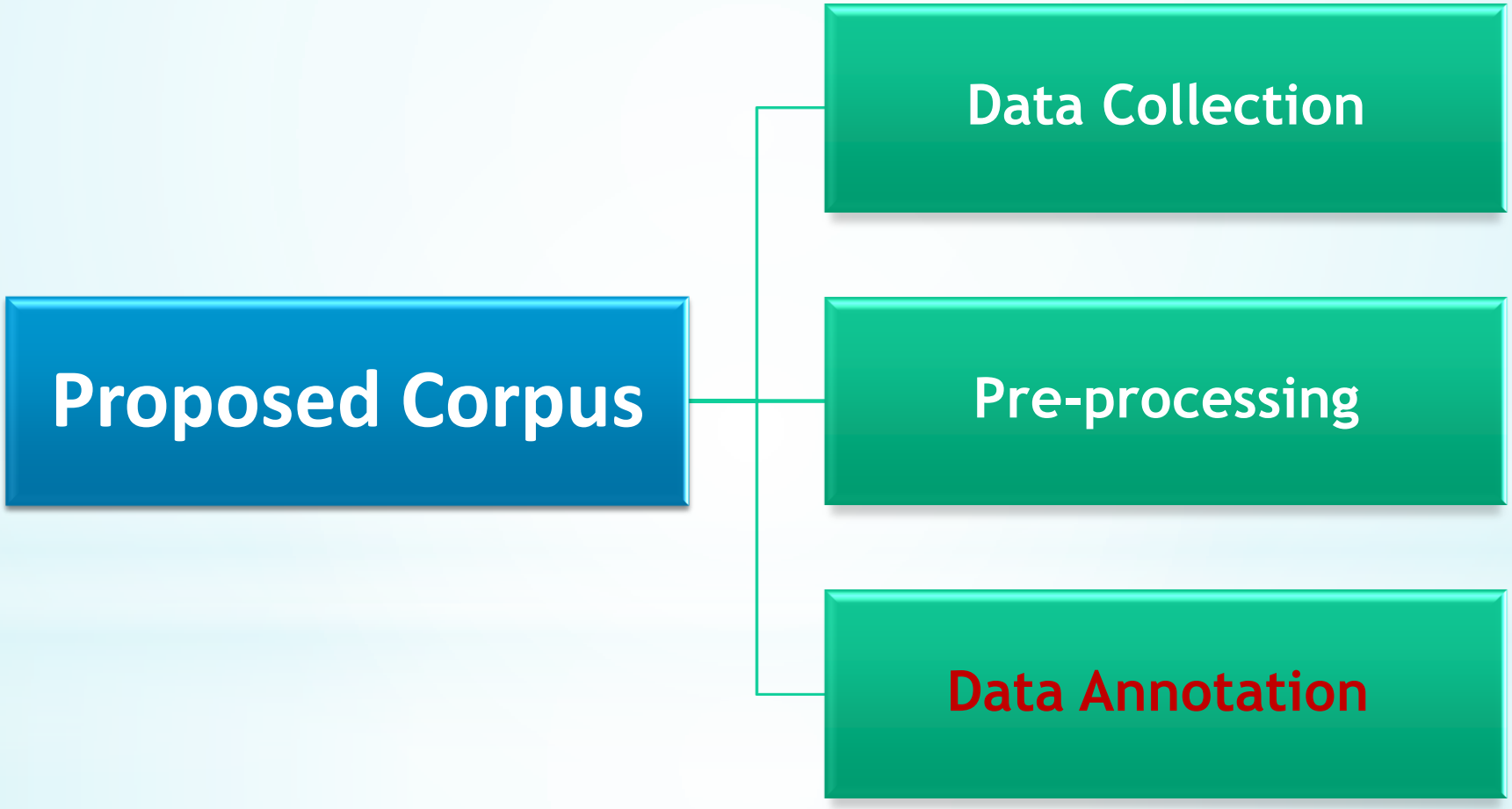
أنت	انـــــت	إنت	انت

All these forms cause the word to be considered as three different words.

Normalization Rules

Rule	Example
Tashkeel	تُقرّب -> تقرب
Repeate	كتيررررر -> كتير
Tatweel	اللـــــــــه -> الله
Hamza	(ء , ىء , ؤ) -> ء
Alef	(آ, أ , إ) -> ا
Lamalef	(لأ , لآ , لا) -> لا
Yeh	(ى , ي) -> ي
Heh	(ة .ه) -> ه

Proposed Corpus (Egyptian Corpus)



Data Annotation

- At this phase, the annotation process is done manually, which manually classify each tweet as *"positive"* or *"negative"*.

Egyptian Corpus

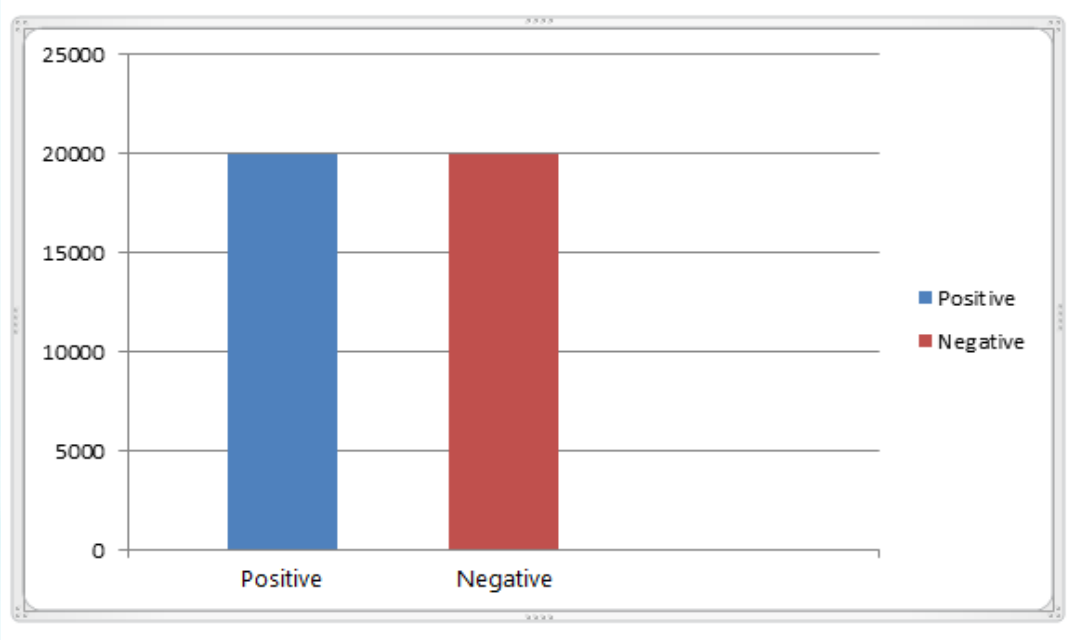
Proposed Corpus (Egyptian Corpus)

- ✓ New Arabic corpus for mixed tweets written in Egyptian dialect and MSA.
- ✓ **Cover several topics including:**

1. Political, Sportive, Health, Social topics.
2. Sarcastic jokes.
3. Proverbs.
4. Poetry.
5. Users' opinions concerning different products.

Proposed Corpus (Egyptian Corpus)

- ✓ This corpus consists of 40000 tweets (classified as 20000 positive and 20000 negative).



Total distribution of Corpus (balanced data)

Proposed Corpus (Egyptian Corpus)

- ✓ It divided into two columns:
  - The first column is called review.
  - The second column is called label.

	B	A
18148	negative	ده اللى انت فهمته . يارينها كانت بركت عليك وهى بتولدك .
18149	negative	بلاش الهجوم الشرش ده .علشان حد اختلف معاك فى الراى .
18150	negative	نيتها اللي ما يموتش من الرصاص يموت مشلول.
18151	positive	اذا كنت تقرا فاني افكر فيك الان و اذا كنت ابتسمت فاني احي بابتسامتك هذه.
18152	negative	فى كمية تفليم وغباء وجهل حاجة ترفع الضغط .
18153	negative	اهم اشخاص فى حياتى .هم ابعد ناس عنى.
18154	negative	يا خرابى على الدماغ.
18155	negative	انا تبت عشان مردتش عليك .
18156	negative	هى افراح العائلة مش بتحلى الا ايام الامتحانات .
18157	negative	للرجيل حكايات بعضها نختارها وبعضها نجبر عليها.
18158	positive	ايدي دي هى اللى هتجيب الملايين بعد كده.
18159	negative	الموافق بتبين مين اللى هيسترجل فى وقت الشدة ومين يمثل .
18160	positive	الله يخليكي، ويجعلها سنة سعيدة مفيهاش صداع في راسنا يا رب.
18161	positive	الحمدلله والله انتى نعمة من ربي.
18162	positive	بصراحه انسانة جدعة جدا ورقيقة جدا ومحترمة جدا وخدومة .انا اتشرفت انى اتعرفت عليكى .
18163	negative	لكل بنت غيرانة مني صدقيني حقك .
18164	positive	انا احتاج شخص يخليني احب الحياه من جديد.
18165	positive	يارب يجي اليوم الي البس فيه كدا .
18166	positive	ده كلام لوجيك جدا .
18167	negative	عايزة اكتب تويته عميقة بس دماغى واقفة .
18168	positive	ربنا يسعدكم ويرزقكم بصبيان وبنات .
18169	positive	يارب انزل على قلبي حب المذاكرة.
18170	negative	الاسعار كل يوم تزيد . حرام بجد .
18171	positive	يعنى تخيلوا حد يخبط عليك ويقولك هابى نيو يير و يديك هديه .معا لترشيح بابا نويل للرئاسه.
18172	positive	اتفضلوا معانا احبتى .
18173	negative	لولا الخيال كنا موتنا من الواقع .

Sample of our corpus

	D	C	B	A
1	Experts#2	Experts #1	label	review
2	N	N	negative	اكبر خطا ترتكبه ان تعامل الناس باخلاقك انت مش باخلاقهم هما .
3	N	N	negative	دائما اكره اخر ليله في كل مكان .
4	N	N	negative	يارب اللي يسرق تويتاتي يدخل النار .
5	N	N	negative	الاسراف فى تناول القهوة يسبب الوفاه .
6	P	P	positive	انا اتبهدلت من التراب النهارده. حاجة تقرف .
7	N	N	negative	فى بنات بتلبس اكسسوارات لدرجة انك تحس ان فى حنطور ماشى جنبك.
8	P	P	positive	احتاج صديق حقيقى يواسينى ويخفف عنى .
9	P	P	positive	لازم اتعلم الثبات الانفعالى زيهم كدا .
10	N	N	negative	جروب الدفعه ليلة الامتحان دايما بتحس ان الناس اللى فيه بتذاكر منهج تانى . وبيكتبوا تعاويذ على الجروب.
11	N	N	negative	النقاش هو الذى يؤدي الى نتيجه اما التعصب فيولد تعصب اكثر منه .
12	N	N	negative	وصلت الى مرحلة انى كرهت بلدى وانى فيها .
13	P	P	positive	عندما قرأت قصه اول شهيدة بالاسلام ام عمار عرفت كم نحن ضعاف .
14	P	P	positive	اجمل انسان هو المتسامح ، المتواضع .
15	N	N	negative	اسمها شبكه تواصل اجتماعيه . مش شبكه مراقبه اجتماعيه.
16	N	N	negative	امي عندها صور لي تنهي مستقبلي.
17	N	N	negative	دنيا ملهاش أمان صحيح .
18	P	P	positive	الرئيس الحنين رزق .
19	N	N	negative	الحكم اللى اتقال النهارده عليهم قاسى جدا .
20	N	N	negative	عند الاستيقاظ تكتئب لمهامك الجديدة .
21	N	N	negative	مصر في المرتبه السابعه في ترتيب الدول الأكثر خطوره في العمل الصحفي انا ابه اللي دخلني اعلام.
22	N	N	negative	سيناريو عبقري وخطة محكمة شفيق ينجح في نفس اليوم اللى مبارك يموت فيه . وشفيق يتجوز سوزان وجمال يورث الحكم.
23	N	N	negative	العرب يسعون لتفسير الاحلام الغرب يسعون لتحقيق الاحلام هذا هو الفرق بين العرب والغرب.
24	N	N	negative	حقيقة غاب صدام فظهر الشيعة .
25	N	N	negative	بعد هذه الاغنية الرائعة لدعم مرسى وهذا الصوت المبهر . اؤكد لكم ان احمد شفيق رئيسا للجمهورية باكتساح.
26	N	N	negative	قد تكون حياتي مليئة بالأصدقاء و لكن قليل ما يشعر بى احد .

Notes:

- we construct the corpus from those tweets having clearly positive or negative sentiment.
- Also, to validate the annotation of corpus, two different experts were asked to check the annotation. The results of both experts were consistent with the corpus annotations with 100 % accuracy.

Egyptian Corpus Statistics

Total number of tweets	40000
Number of positive tweets	20000
Number of negative tweets	20000
Number of words	359,818
Max tweet token	39
Number of tokens	1,953,869
Average tokens per tweet	17

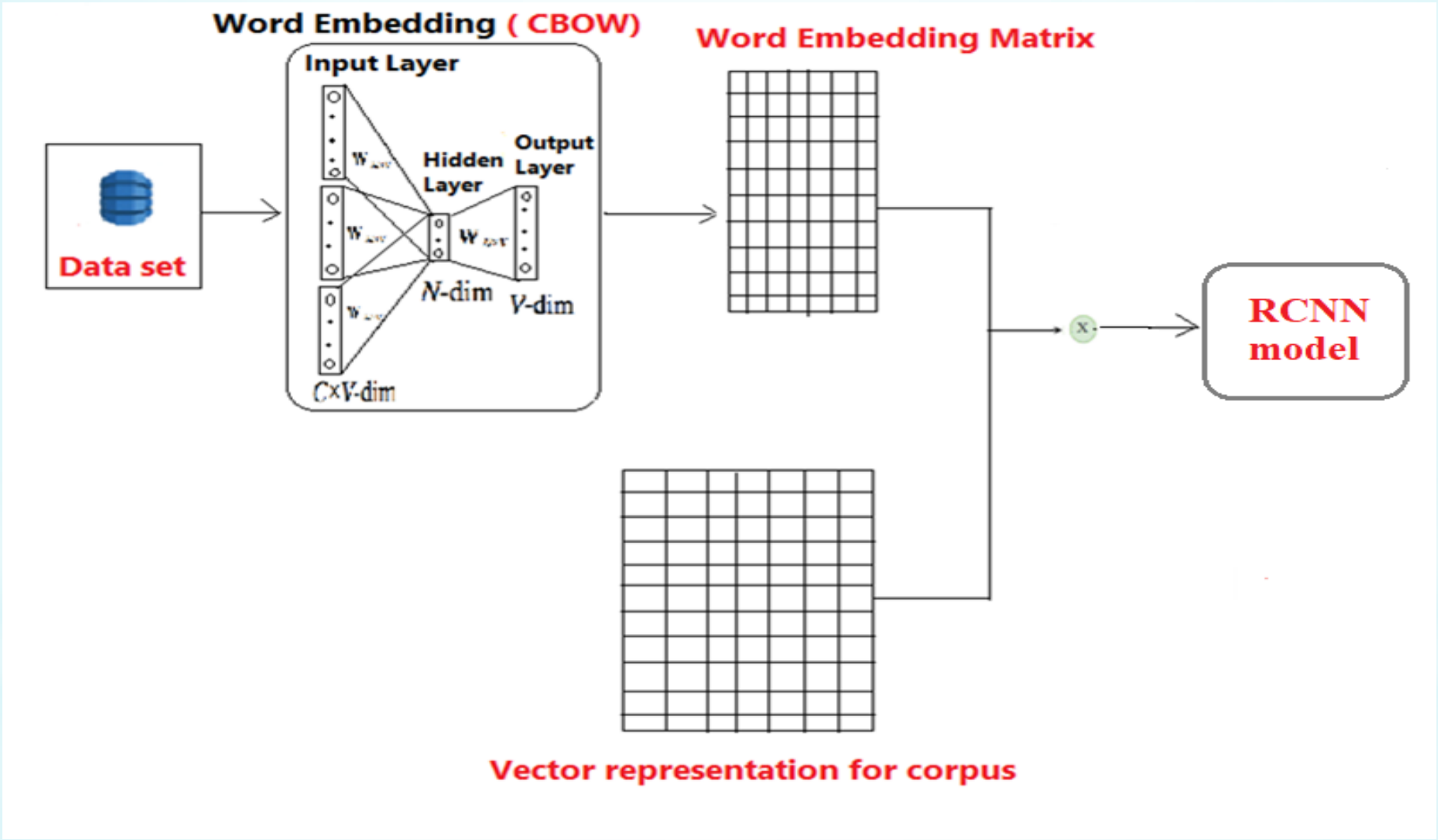
✓This corpus is available in [6].

6- Kora, R. and A. Mohammed, "Arabic tweets Egyptian dialect", . Mendeley Data, V1, 2019. *https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LBXV9O&version=DRAFT&faces-redirect=true*

Agenda

1. Introduction.
2. Proposed Corpus.
3. Embedding Layer.
4. Architecture of RCNN Model.
5. Results and Discussion.
6. Publications.
7. RCNN Code.

Embedding Layer of RCNN Model



Agenda

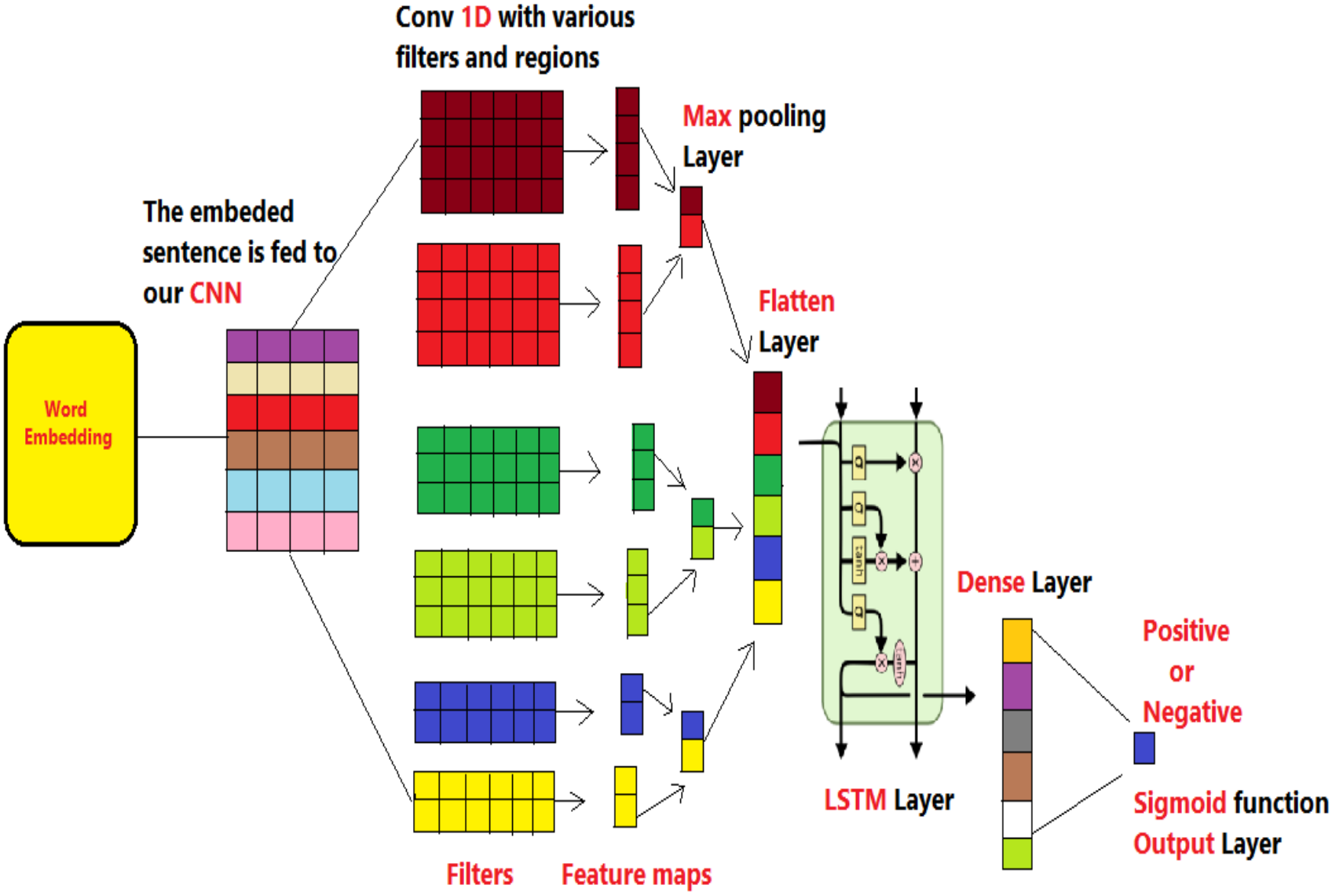
1. Introduction.
2. Proposed Corpus.
3. Embedding Layer.
4. Architecture of RCNN Model.
5. RCNN Code.
6. Publications.

RCNN Model

- Recurrent Convolutional Neural Network (RCNN) is a network the combining the advantages of both CNN and LSTM neural networks.

- The RCNN architecture consists of: Embedding Layer, Convolutional Layer, Max-pooling Layer, Flatten Layer, LSTM layer, and Dense an activation layer.

Architecture of RCNN Model



Agenda

1. Introduction.
2. Objectives.
3. Proposed Corpus.
4. Proposed Approaches.
5. Publications.
6. RCNN Code.

Agenda

1. Introduction.
2. Proposed Corpus.
3. Embedding Layer.
4. Architecture of RCNN Model.
5. Publications.
6. RCNN Code.

Publications

1. *Mohammed, A., Kora, R.: Deep learning approaches for arabic sentiment analysis. Social Network Analysis and Mining 9 (1), 52 (2019)* Springer. *DOI 10.1007/s13278-019-0596-4. URL* https://link.springer.com/article/10.1007%2Fs13278-019-0596-4

Social Network Analysis and Mining (2019) 9:52
https://doi.org/10.1007/s13278-019-0596-4

ORIGINAL ARTICLE

Check for updates

Deep learning approaches for Arabic sentiment analysis

Ammar Mohammed[1] · Rania Kora[1]

Received: 27 June 2019 / Revised: 3 August 2019 / Accepted: 9 September 2019

Abstract
Social media are considered an excellent source of information and can provide opinions, thoughts and insights toward various important topics. Sentiment analysis becomes a hot topic in research due to its importance in making decisions based on opinions derived from analyzing the user's contents on social media. Although the Arabic language is one of the widely spoken languages used for content sharing across the social media, the sentiment analysis on Arabic contents is limited due to several challenges including the morphological structures of the language, the varieties of dialects and the lack of the appropriate corpora. Hence, the rapid increase in research in Arabic sentiment analysis is grown slowly in contrast to other languages such as English. The contribution of this paper is twofold: First, we introduce a corpus of forty thousand labeled Arabic tweets spanning several topics. Second, we present three deep learning models, namely CNN, LSTM and RCNN,

2. *Kora, R., Mohammed, A.: Corpus on Arabic Egyptian tweets (2019). DOI 10.7910/DVN/LBXV9O. URL https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LBXV9O&version=DRAFT&faces-redirect=true*

Agenda

1. Introduction.
2. Proposed Corpus.
3. Embedding Layer.
4. Architecture of RCNN Model.
5. Publications.
6. RCNN Code.

thank you