## PROJECT PHASE -1

THİS DOCUMENT WILL CONTINOUSLY BE UPDATED DURING THE PHASE-1. Use Revision

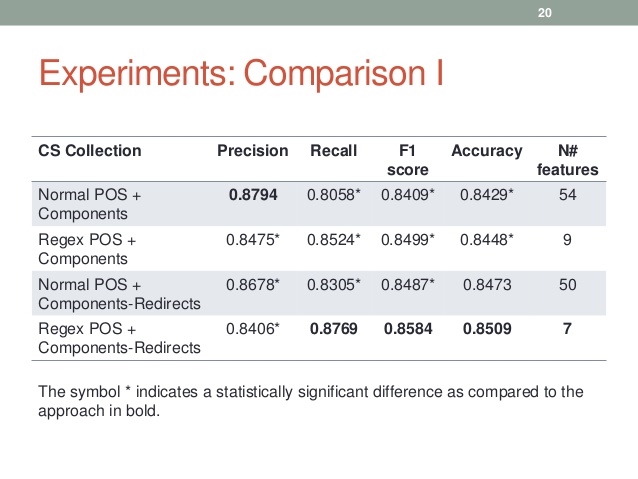
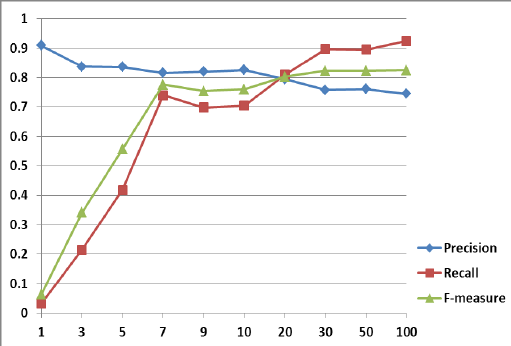
DATE-1

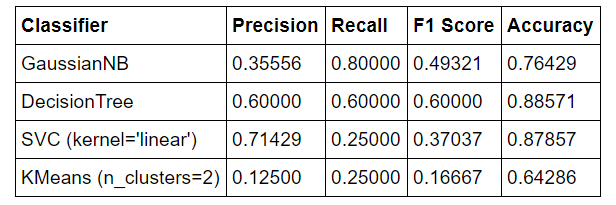
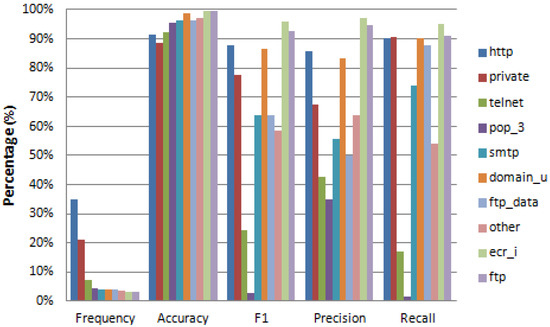
START DATE 14.12.2017

Phase-1 must be finished before 19 dec.

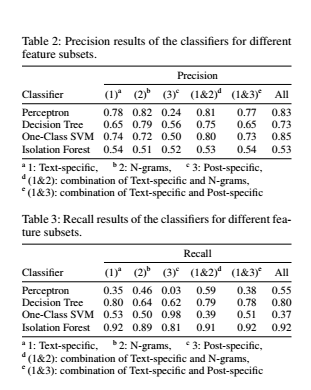
\*\*\*\*\*\* start of updated section\*\*\*\*\*\*

1. Could we add accuracy, F1, specifity to the result showing? And the app should show some jeneric results and plots if possible. Some examples are shown below. These are only to give ideas. Especially for the different domains sport, politics the scores change sharply? Even if we have a small dataset now but we will have some bigger by te help pf the app itself and we will be able to see that difference so it is very important to think that way.

[](https://www.google.com.tr/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwjQ4OXE85DYAhUII-wKHf6MBJIQjRwIBw&url=https%3A%2F%2Fwww.slideshare.net%2FeXascaleInfolab%2Feffective-named-entity-recognition-for-idiosyncratic-web-collections&psig=AOvVaw1DhjXxbXC_Nx0a7ykoID-T&ust=1513593359534859)[](https://www.google.com.tr/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwiVpYmx8JDYAhUSpKQKHXGwDtgQjRwIBw&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2F275523467_fig2_Figure-5-Trend-lines-of-Recall-Precision-F-measure-for-NFR-classification-on-1000&psig=AOvVaw1DhjXxbXC_Nx0a7ykoID-T&ust=1513593359534859)

[](https://www.google.com.tr/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=0ahUKEwjpuNev9JDYAhVNUlAKHWH5DN4QjRwIBw&url=https%3A%2F%2Fmedium.com%2F%40williamkoehrsen%2Fmachine-learning-with-python-on-the-enron-dataset-8d71015be26d&psig=AOvVaw1DhjXxbXC_Nx0a7ykoID-T&ust=1513593359534859)

1. I am not able to run the app yet but as i see we use tf-idf as a base feature. And then n-grams. The table below shown belown is some results of the classifiers for different feature subsets. The ultimate result that i want to reach for the **main algortihm** will be something similar to this.

[](https://i.stack.imgur.com/enEIt.png)

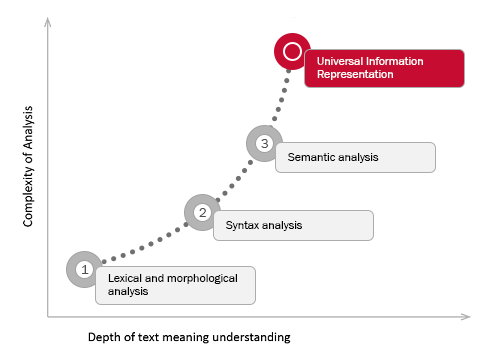
And an interface with a selectional parameters (for feautures) and classifiers will be apart from the main training algorithm.

1. Could you add some features shown below to make the project orijinal. For example;

* Using sytle markers like:
* The average sentences number
* The number of slang words per text
* The number of punctuation marks (conjuctions, exclamations, points, comas etc.)
* Number of inverted sentences (this is some related to PCFG) CKY algorithm should be read to understand this.

|  |
| --- |
| You can use the slang dictionary of <https://www.turkedebiyati.org/argo-sozlugu.html>  For example in that address it shown as below:  **Abondone:** pes etmek **abanmak:** birine yük olarak onun sırtından geçinmeye bakmak **abtestini vermek:** azarlamak   * **As you can estimate first word, then the explaination comes after colon/double dot. Also it is important search them before stemming parcing. Before parsing or preprocessing the sentences you can firstsearch for this words or words group. For example if you stem abanmak**   **You can find aba as the root so you may skip the slang by this way** |

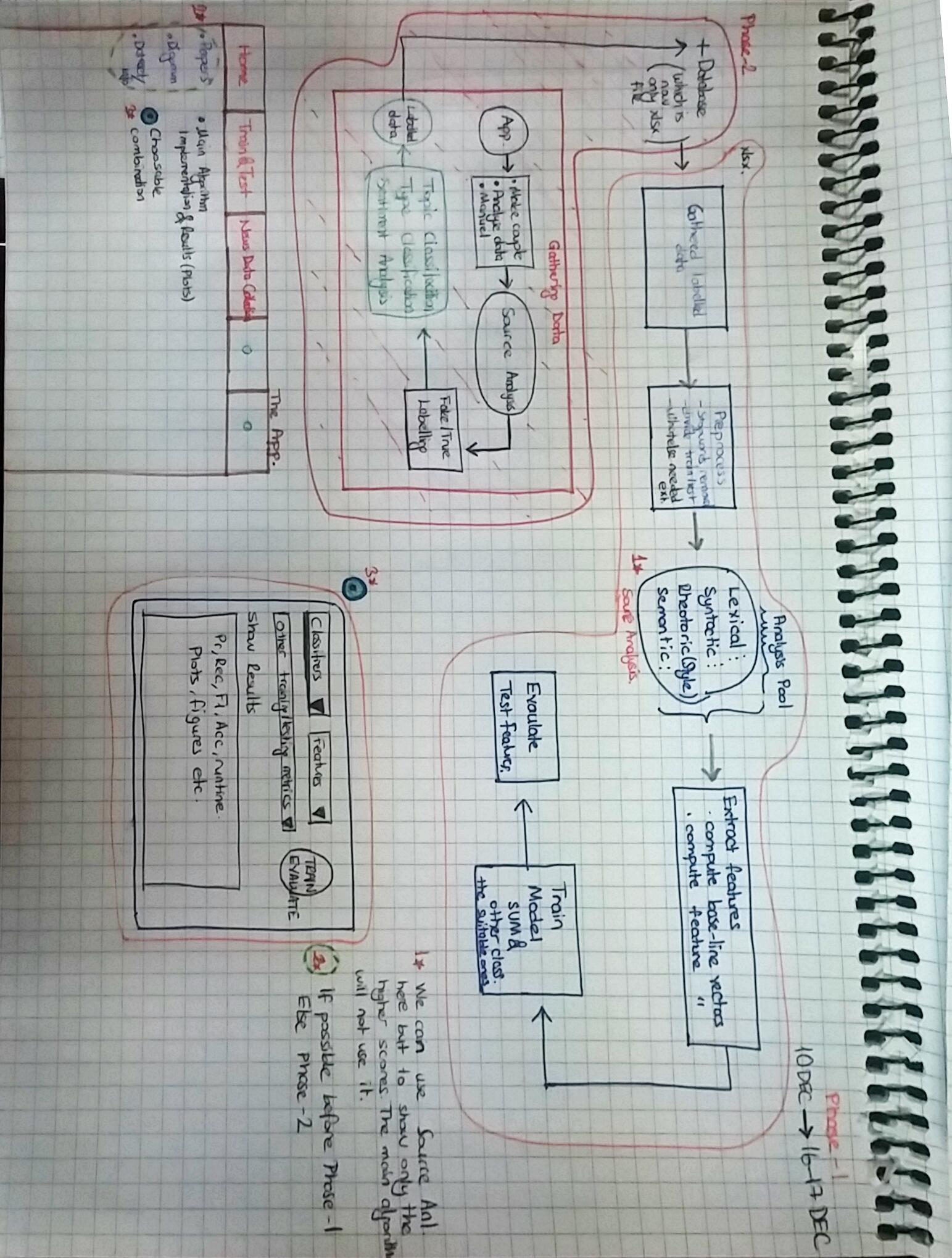
1. The use of some suffix (appendix) may be an indication of some fake or true news so may be it will be useful to make use of n-grams (2,3) for the letters (for example i observe sa, se, mış, miş, muş, müş etc are mostly used in fake, and (di, dı, du, dü, tı, tü, tu, ti etc. Are mostly in true
2. And a normalization (foreign keys, unique ids to reach faster) may be needed for the database when the database gets bigger. We dont use any database here we use the files. We have to keep in my mind about that.
3. Also POS (Part of Speech Tagger) colud also be used.



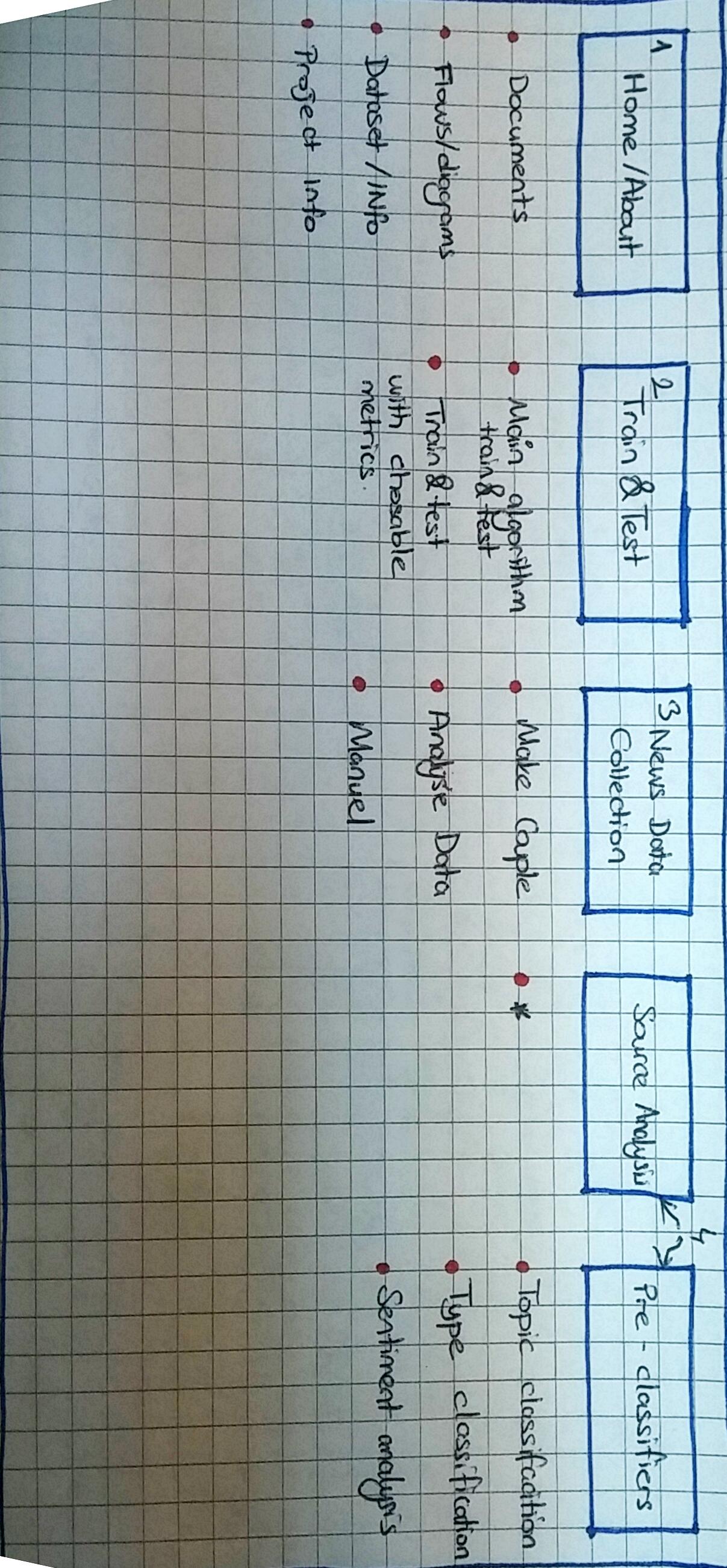
We may make use of each analysis section in spesific and together.

\*\*\*\*\*\* end of updated section\*\*\*\*\*\*

SUMMARY DRAWING OF PROJECT DRAFT



The main algorithm have to be very robust and original.



**Short Description for Phase-1**

To set a model of algorithm for detecting Turkish news whether it is true or fake

* A desktop app providing

1. **Home Section** (Papers, diagrams, dataset, info/about)

*There will be some subsections which will be filled by some information about the project as PAPERS, BLOCK DIAGRAMS, DATASET, PROJECT INFO. If we can finish by time we can fullfill these but these are the priorities.*

|  |  |
| --- | --- |
| Documents | From here the documents (pdf, word, excel etc.) can be added, deleted, updated. And shown to the user. |
| Flows/Diagrams:  Dataset/Info |  |
| Project Info |  |

1. ***Train&Test (Implementation Section)***

(Pages showing results, statistics, plots)

Main Alg.: Running the project and giving the user what is done for information simultaneously.

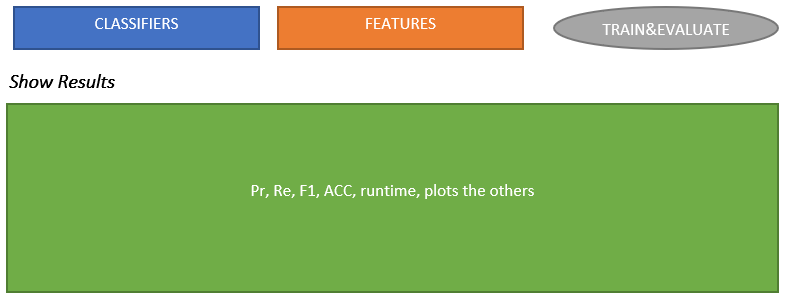
Train&Test: With chosable metrics

*Main algorithm, trained model will be here to implement. And;*

*This will be a section of implementing the algorithm. You way use this as a trigger mechanism for user. It is better to show something to last user than saying this program works ……. And these are the results. It is up to your imagination.*

*Fixed main algorithm obtained best results. We have to compare with some as a sample shown Sample-1 AT NEXT PAGES.*

*What I want here is apart from the best methods (algorithm we found by training), features or combined feature (figure-1). We will be able to do some experiments with preferable features, classifiers and basic combinations and store the results in case of repetable use (it is not needed to repeat every step with the same choices and same database)*

**

*Figure-1*

1. News Data Collection

*In PHASE-1 we will concentrate in this section. The requirements ans spesifications for this section is shown at* **APP MOCK-UPS Section-7.**

Make Couple:

Analyse Data:

Manuel:

1. Source Analysis
2. Pre classifiers

Topic classification

Type Classification

Sentiment Analysis

1. **The development environment:**

Python with many packages is preferrable for doing this but we may make use of R prog to see eye-pleasing plots app.

* NLP packages
* Zemberek or Snowball stemmer for Turkish for tokenization (or some else if there is.)
* The usage of all packages except for defaults must be stated.
* If it is useable for windows the other addings colud be made.

***2.1 Classifiers:***

***FAKE & TRUTH Classification***

SVM (more suitable for this kind of stuff) and one or two of the other most accurate classifiers (some examples for it shown below) willl be chosen make model comparable between the classifiers.

* Naive Bayes, Stochastic Gradient Descent(SGD), Gradient Boosting, Gradient Boosting, Bounded Decision Trees, Logistic Regression, Decision Trees

Note: In the app it should be chosable for different classifiers but main algorithm is for the best result. Now focus on only the main algorithm (Phase-1

***NEWS TOPIC Classification*** (PHASE-2 /Will be detailed)

This app will also be able to classify the topics of news

***Sentiment Analysis*** (PHASE-2 /Will be detailed)

In the dataset the column sentiment is provided. So the algorithm has to find the the sentiment value of the news context(text).

***SOURCE Analysis*** (PHASE-1 MiLESTONE-2 /Will be detailed)

--- You said it cant be done by 17 DEC so we can carry these into Phase-2 even it brokes my planned framework. Because unfortunately I have to go on manuelly labelling till it is done.

My insist on finishing that in Phase-1 is only for getting the data bigger and we can obtain some better results. By the methods we will use in it I mean the recommendation system provided by Source Analysis and subsequently 3 classfications which are as you said very important points will make the project differs from the litearature.

As shown in the block diagram, our system has to make evaluation and after that presents the user/annotator the findings on as a recommendation with the Source Analysis results.

If the user/annotator wants it to label manually or not.

This does not mean do not use source analysis for deciding fake or true value it can also be shown but the characteristic of our dataset is intensively related to few sites which is already labelled for delivering verified or satiric news then it will be only self-deception to detect by using source analysis. Our first aim here is using source analysis to make a qualified dataset getting bigger day by day. You will understand what i mean while reading the rest.

***Type Classification*** (PHASE-2 /Will be detailed)

*Q-3 The algoritm we will use for classifying in phase-2 is a bit more free than PHASE-1 So ı am open to your offers if you have*

***2.2 Comments and Description:***

Just for a reminder: A detailed methodology description for codes and project itself willl be needed. So at the end of the PHASE-1 MILESTONE-1 there will be description especially for how to find and use the feature, the methods and algorithm, plots and the usage of program should easily be understood. A short education which we can make together after the PHASE-1 M-1 completed.

***NLP for Turkish Language:***

As in all languages there are rules and spesific characteristics for Turkish language. I will try best for you to comprehend. In this part i ll always update the document whenever I found something declarative for you. For a start point i recommend you to look that link. Even it is written in Turkish youcan easily understand from the sample codes.

<http://www.veridefteri.com/2017/11/20/turkce-metin-islemede-ilk-adimlar/>

1. **Dataset:**

Our dataset is close to thats’ shown below

<https://www.researchgate.net/profile/Victoria_Rubin/publication/301650504_Fake_News_or_Truth_Using_Satirical_Cues_to_Detect_Potentially_Misleading_News/links/571fed4c08aeaced788acd8e/Fake-News-or-Truth-Using-Satirical-Cues-to-Detect-Potentially-Misleading-News.pdf>

*First Set (Verificated News)*

Source: turkish news verification platforms (mostly from <https://teyit.org/> and a few others & correspondence of the news from other news sites) to from a FAKE-TRUE pair

|  |  |  |
| --- | --- | --- |
|  | From other news sites | Mostly teyit.org |
| TOPIC | FAKE | TRUE |
| Urban Legend | 10 | 10 |
| Sports | 10 | 10 |
| Politics | 10 | 10 |
| Health | 10 | 10 |
| Others (Science, Environment, Social) | 10 | 10 |

Note:

This set will be gathered automatically by web crawler at Phase-2 but now unfortunatley manuel.

*Second Set (Satirical News)*

Source: turkish satirical news platforms (<http://www.zaytung.com/> and a few others & correspondence of the news from other news sites) to from a FAKE-TRUE pair. But some of the news (very few) hasn’t got true or fake correspondence (at most ten, this is needed make algorithm spesific)

|  |  |  |
| --- | --- | --- |
|  | From other news sites | Mostly zaytung.com |
| TOPIC | TRUE | FAKE |
| Urban Legend | 10 | 10 |
| Sports | 10 | 10 |
| Politics | 10 | 10 |
| Health | 10 | 10 |
| Others (Science, Environment, Social) | 10 | 10 |

*Third Set (Common Related to Phase-2)*

*Columns*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *NewsId* | *To reach the data faster. It can be useful for us to store data in DB after the PHASE-1 when our data gets bigger (unique identity).* | | | |
| *DatasSetId* | *Three set of data detailed in above* | | | |
| *NewsDate* | *The date of the news or when archieved. Some of them are absent espacially fake ones)* | | | |
| *Actors* | *Proper nouns, starting words and subjects (we may change this into the metadata labelled in site for header.)* | | | |
| *Sentiment* | *The feeling (sentiment of the author* | *Positive* | *Negative* | *Objective* |
| *Type* | *RUMOUR (RUM)* | | | |
| *INFORMATION&CLAIM (INF\_CLA)* | | | |
| *TRUTH VERIFICATION (TRU\_VER)* | | | |
| *DISINFORMATION&MISINFORMATION (MIS\_INF)* | | | |
| *HUMOUR (HUM)* | | | |
| *PROPAGANDA&MANUPULATION (MAN\_PROP)* | | | |
| *Value* | *Fake or Truth* | | | |
| *FakeTrueMatchNewsId* | *The correspondence(NewsId) if there is one. If no then 0* | | | |
| *Topic* | *URBAN LEGEND* | | | |
| *POLITICS* | | | |
| *HEALTH* | | | |
| *SPORTS* | | | |
| *OTHERS (Science, environment, social etc)- WHICH can not seperable from others easily. Data driven observation* | | | |
| *URL* | *Url or social media virals* | | | |
| *News Title* | *Header* | | | |
| *News text* | *Context, whole text (min 3 sentences max 16 sentences)* | | | |
| *GlobalEventId* | *GDELT project relatedness (PHASE-2)* | | | |
| *Gdelt Factor* | *GDELT project relatedness (PHASE-2)* | | | |

*Some small others can be added during phase -2 And also i will fix some badness of tha data during all the phases.*

1. **Preliminary Work:**

As in every NLP project our algorithm the data (the text) is needed to be formatted, tokenized and vectorized (Getting rid of stop words ext). Vectors and matrices formats for data will be obtained that way.

Turkish is an agglutinaive language so it is important for the originality of our study; stemming will needed. For ex. Human (as in turkish insan) “insan, insanlığa, insanlık, insanlığa) belongs to the root “insan” so getting all the words as they differ makes our feature space bigger and the spesificness never will be gained for the features. So we have to apply [Zemberek](https://code.google.com/p/zemberek/) or [Snowball Stemmers](https://pypi.python.org/pypi/snowballstemmer) (most known Turkish stemmers/morphologic analyzers). And also this will help us to correct the wrongly written words. There are good examples of usage for zemberek is shown

<http://www.b4deploy.com/sorular/zemberek-example-for-python/>

<http://gurmezin.com/python-ile-zemberek-kutuphanesini-kullanmak/>

* If it is possible as a future ability to use both of them (Z and S) compare will be very good.

*Q-4 Do you have any offer about the opinion above?*

1. **Methodology:**

For starting point it will be trained by the dataset. So we can call as it is a supervised learning. Now manuelly labelling is going on by some experts. The sample data contains 56 news (%99 from 2017) right now but it will be by 72 11 DEC and will be bigger day by day till 13 DEC.

Also our model will be able to use a web crawler to gather data So by the help of our web crawler the dataset can easily get bigger (Will be explained ...).

* Lexical and Morphological, Syntax, Semantic and Source Analysis will be used in this project.
* Data Driven-Linguistic Analysis (both for feature extraction or creating feature) (will be explained)

*Q-5 I wanted to use Semantic Analysis but can not figure out that can you help me about that*

This will inform you about the analysis models and the literature.

<https://www.researchgate.net/publication/281818865_Automatic_Deception_Detection_Methods_for_Finding_Fake_News>

There are lots features for this models.

I will try to explain much of them to use. Our algorithm make use of most effective features or combination of features and also try to New ones. Most important part of the project is FEATURE SELECTION and EXTRACTION.

You will be given some hints (at the section …) about the data so by these hints we can try to find New features for the originality. This paper in link is a good example of extracting New features.

<https://www.researchgate.net/profile/Victoria_Rubin/publication/301650504_Fake_News_or_Truth_Using_Satirical_Cues_to_Detect_Potentially_Misleading_News/links/571fed4c08aeaced788acd8e/Fake-News-or-Truth-Using-Satirical-Cues-to-Detect-Potentially-Misleading-News.pdf>

I estimate that this feature fit the case more sutiable but in here

TF-IDF

n-GRAM of words (1-2)

– in some studies they do it also for letters (I mean the fake news writig style sometimes use certain (addition, suffix) I ll try to give god examples as soon as possible.

Style Markers (This feature can also thought one by one or a normalized combinational feature)

* The number of slang words or usage per text (a slang vocubulary will be given to you)

You can use the slang dictionary of <https://www.turkedebiyati.org/argo-sozlugu.html>

For example in that address it shown as below:

**Abondone:** pes etmek  
**abanmak:** birine yük olarak onun sırtından geçinmeye bakmak  
**abtestini vermek:** azarlamak

* **As you can estimate first word, then the explaination comes after colon/double dot. Also it is important search them before stemming parcing. Before parsing or preprocessing the sentences you can firstsearch for this words or words group. For example if you stem abanmak**

**You can find aba as the root so you may skip the slang by this way**

* Number of certain punctuations (conjuctions, exclamation, points, comas etc)
* Words number
* Sentences number
* Average words number in sentences
* Average sentences number in text
* Number of inverted sentences of all sentences
* Number of incomplete of all the sentences
* Numbers of verb in a sentences
* Number of adverb in a sentences
* Number of adj in a sentences
* Number of particle in a sentences

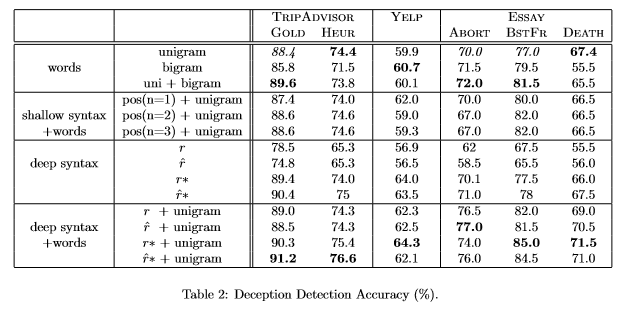
One of the studies will be very helpful for the characteristic of turkish. But it is in turkish so i will try to find of its english. But have a look at it to understand.

<https://www.researchgate.net/publication/319910626_Turkce%27nin_Karakteristik_Ozelliklerinden_Yararlanarak_Yazar_Tespit_Etme_Taking_Advantage_of_Turkish_Characteristic_Features_to_Achieve_Authorship_Attribution_Problems_for_Turkish?enrichId=rgreq-b847dd772d29bb748d23f4f4575c5d2b-XXX&enrichSource=Y292ZXJQYWdlOzMxOTkxMDYyNjtBUzo1NDAzMDM3NDY0OTQ0NjRAMTUwNTgyOTg1OTQ4Mg%3D%3D&el=1_x_2&_esc=publicationCoverPdf>

At the last we will try to reach such kind of results tables (Only to give opinion not exactly the same) <http://www.aclweb.org/anthology/P12-2034>

We can modify this add/minus some (for example we can add Style Markers, the data-driven features or the extracted feature)

*Q-6 Something that i am trying to figure out is data driven observation features so if I can I will share with you this makes the results better. Do you have any opinion or advice?*



**Sample-1**

* Topic Classification Accuracy
* Sentiment Analysis Accuracy
* Type Clasification Accuracy

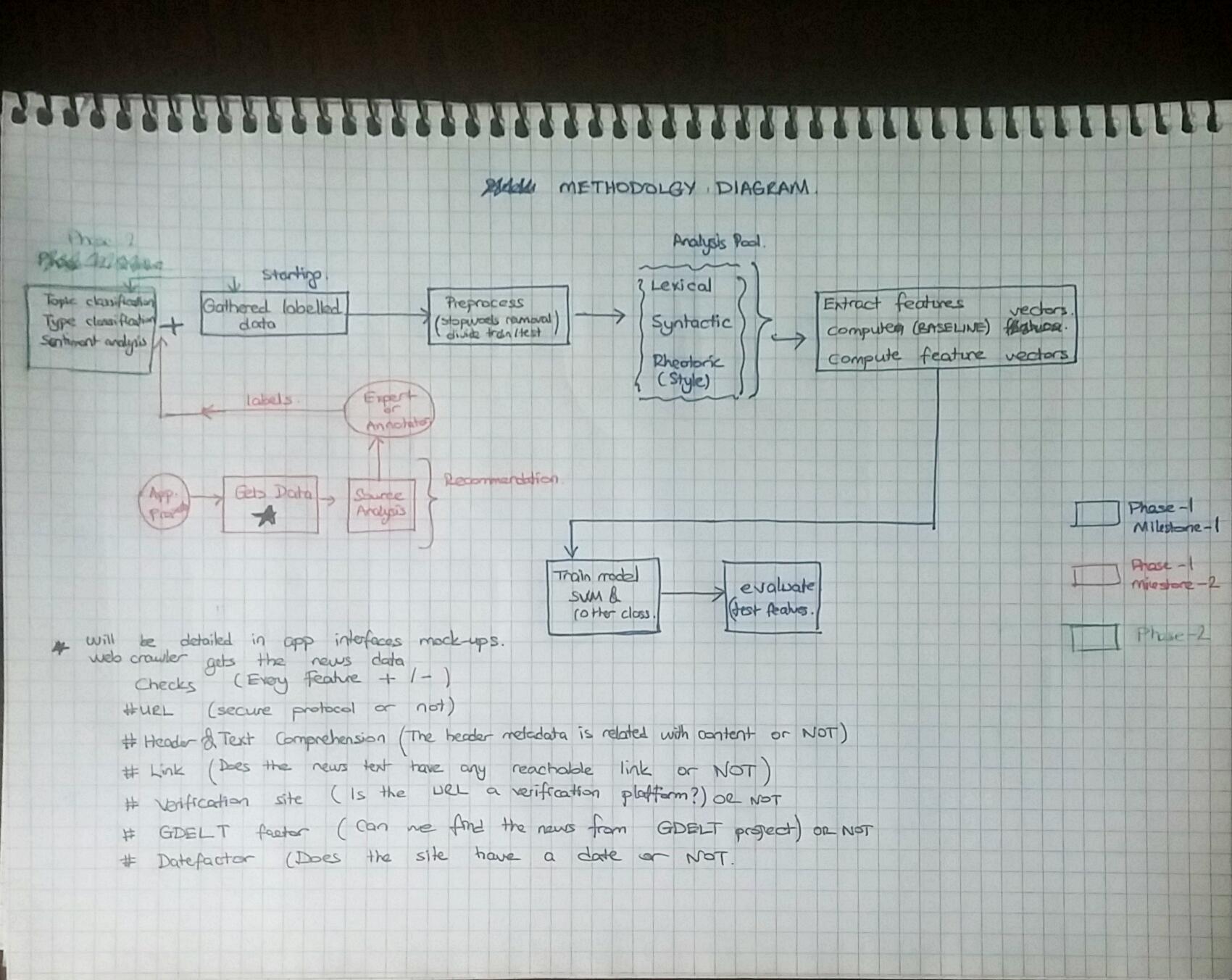
Your starting point could be this links

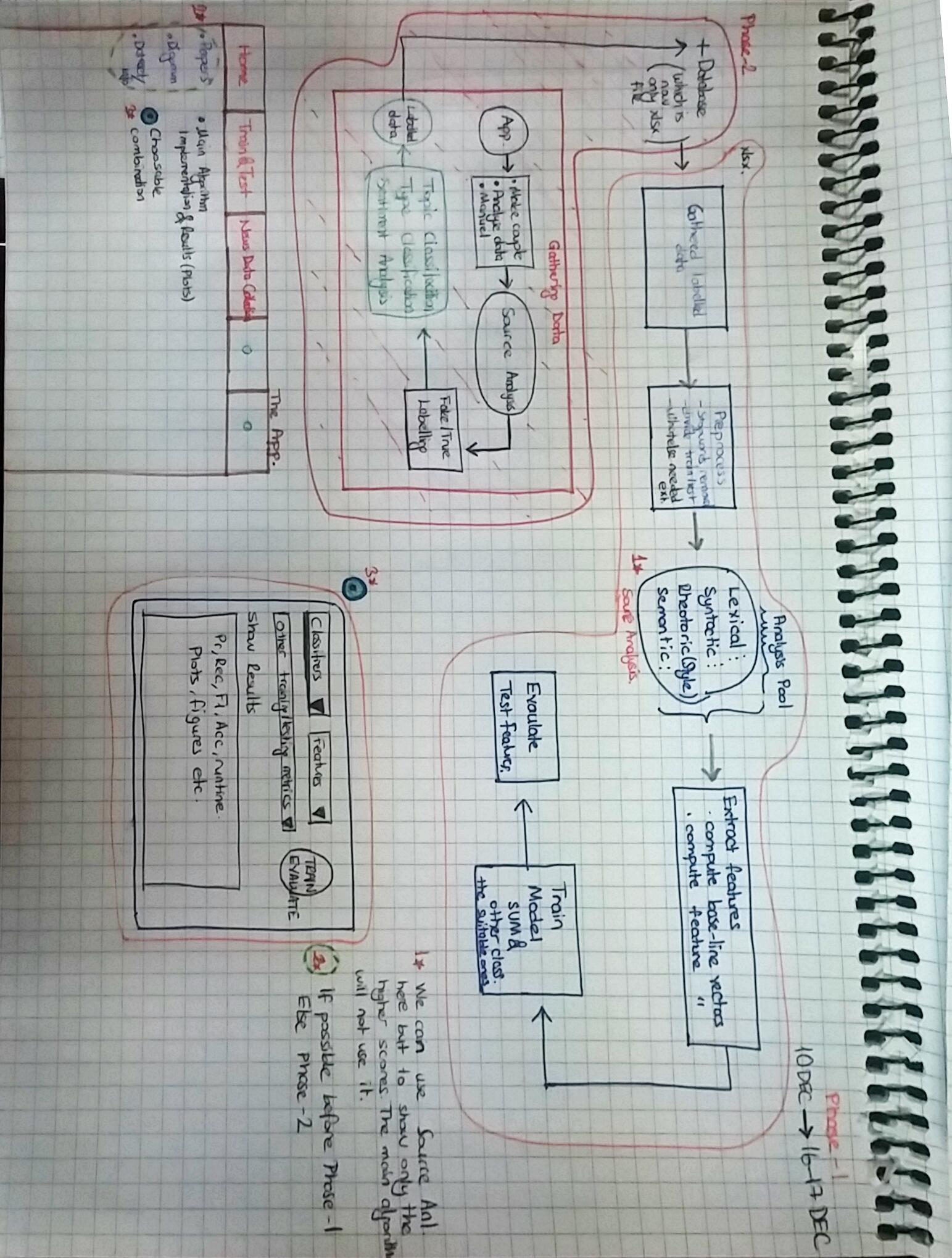
<https://github.com/aldengolab/fake-news-detection>

<https://github.com/narain280493/Deception-Detection>

But please do not copy all because turnitin or such kind of identification tools will be tried it must be original.

1. **Old and Updated Blok Diagram:**

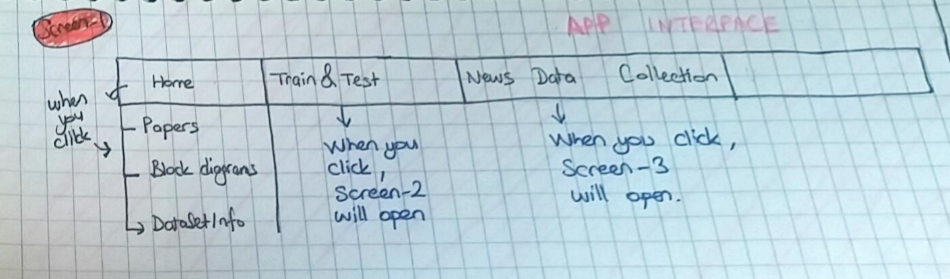
****



1. **APP MOCK-UPS**

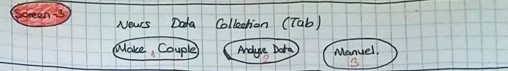
These are all pencil drawing because we dont have enough time. So excuses for that. You can also change the name labelling of tabs, buttons whatever i only tried to make their goals clear in here

In Screen-1 shown below is the main interface of the app. You are limited to this you can make it on your own decision but some important points that i want to emphasise about Screen-3.



Screen-1

When we click Screen-1 News Data Collection SubSection. This will show us Screen-3 that is responsible for how to gather the data and also provides to implement these methods.



Screen-3

This section is gathering data by web-crawler and manuelly. It will be using Source Analysis. We avoid from using whole source analysis columns

**Make Couple:**

In this section for our verification based set (LABELLED V). In this section we will give the

Spesific URL to the system and SEARCH button will find the news. And Header, topic and the text will be shown on the related places at the interface.

We have to decide max sentences number to get (30 sentences)

For example. We gave the system the URL =

<https://teyit.org/toplu-igne-kullanarak-kalp-krizi-geciren-kisinin-hayatinin-kurtarilabilecegi-iddiasi/>

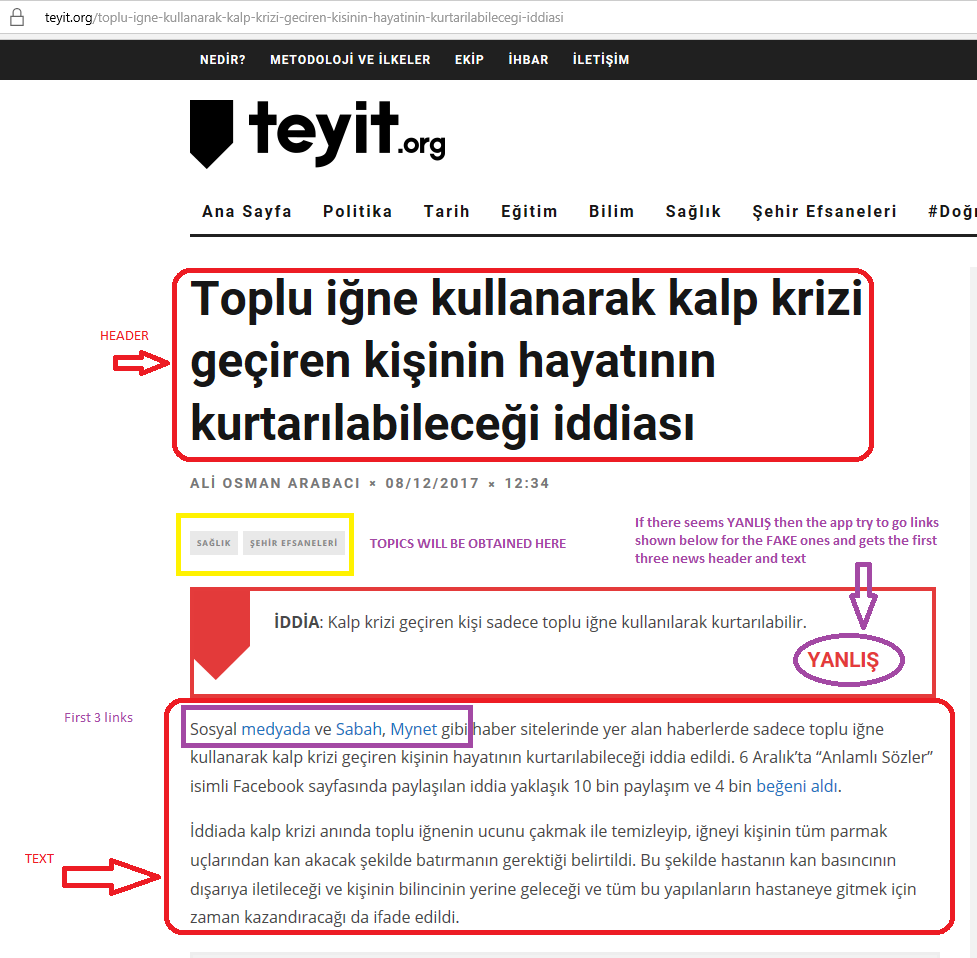
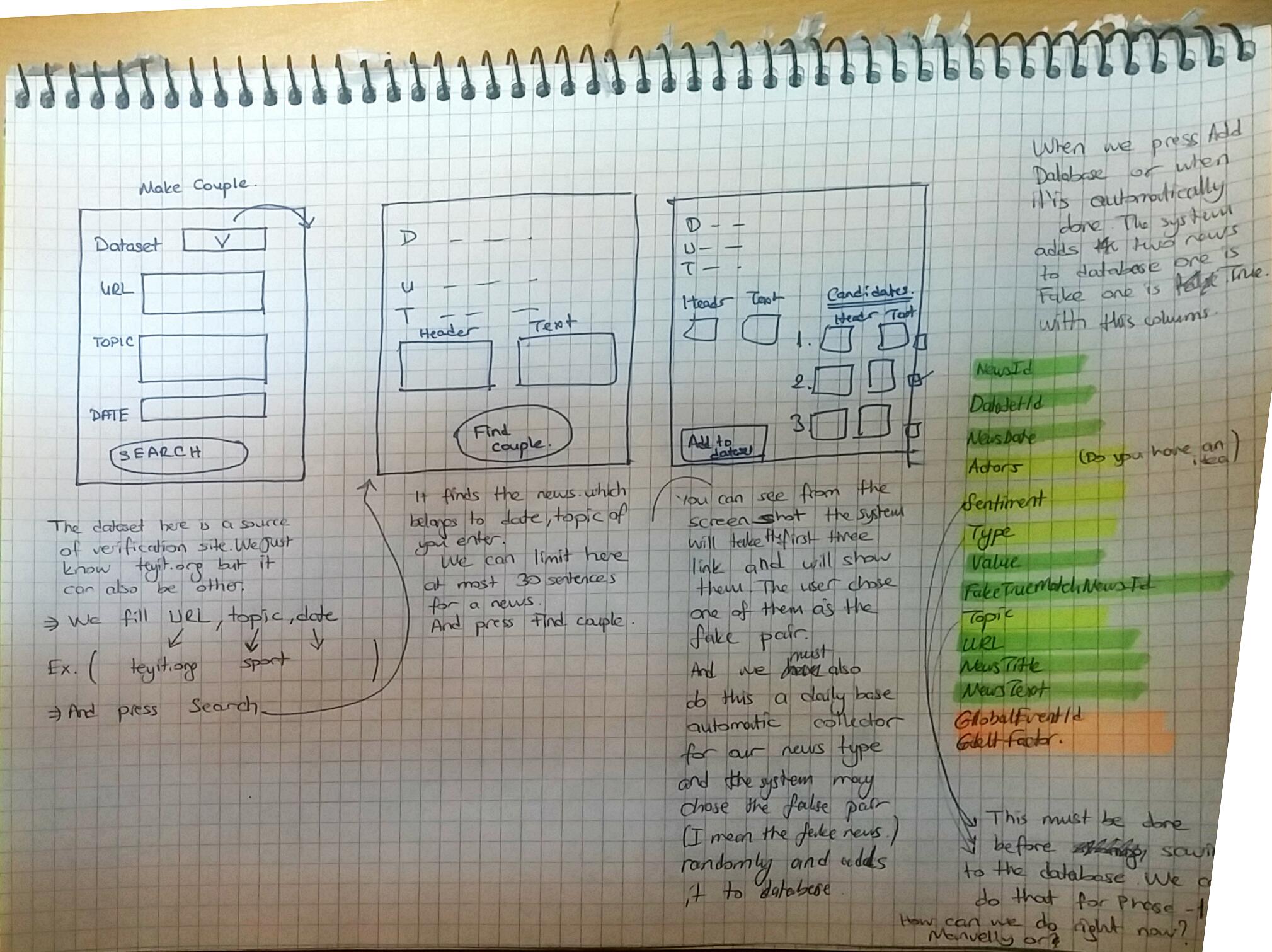


Image-1

And the webcrawler will go and look for the site (IMAGE-1). Then the possible couples of the news is shown belown in a such an interfaces when press find couples. The user decides the pair with checking which he/she wants. If the user wants to add the news into database with the chosen pair then a true-fake couple will be add to database.



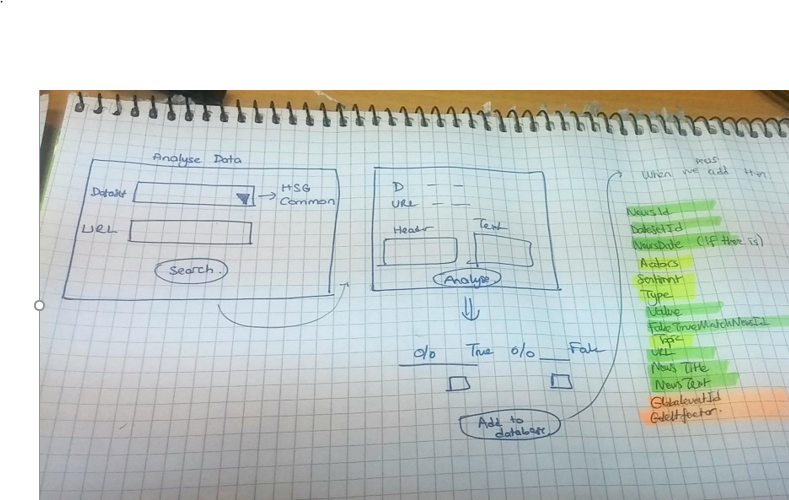
For Phase-2 or may be Phase-3 a daily base automatic web crawler system will add these to the database. One problem here is the other columns must be labelled manuelly this is why we have to label them manuelly till Phase-2. That is we use source factor as a recommender because we all know about source itself before. Sorry above Topic is already labelled so it is not needed. Type, Actors and sentiment classification is needed (yellow lighted)

**Analyse Data:**

We will give URL again same as Make Couple but here we label the set of data HSG, V or Common. Till now for not to spoil the data we dont use any Common labelled data. We dont try to make a couple here. We only try here for whether a unique news is fake or not. And then a DIRECTLY PROPORTIONAL method (which is not algorithmically I NEED help here may be a normalization can be made here) will be used by the app that uses the properties to below to recommend user for the news is true or fake. For ex. If a news has 4 yes and 3 no then the news it means 4/7\*100 % true and 3/7\*100 % fake for recommendation. If the user wants to add the news into database with the chosen

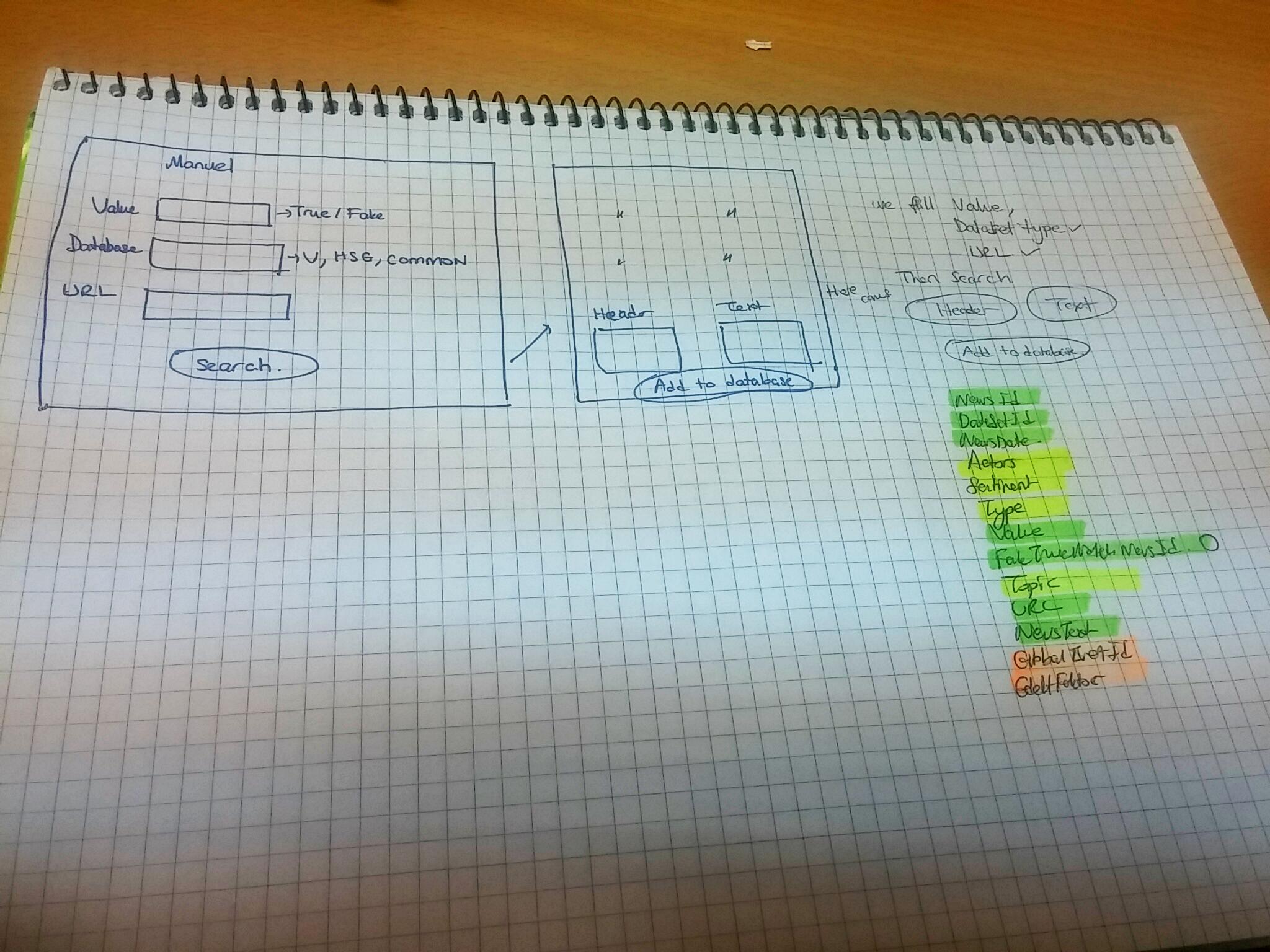
|  |  |  |
| --- | --- | --- |
|  |  |  |
| URL CHECK (Secure or not) https /http | YES | NO |
| Header & Text Comprehension | YES | NO |
| Links (Reachable) | YES | NO |
| Is this a verification site? | YES | NO |
| GDELT factor (if it is foundable from that project database) | YES | NO |
| Date factor | YES | NO |
| Slang word dict. Search | YES | NO |
| Is this site a satirical/fantastic news site? | YES | NO |

Some other observation driven features can be added.



**Manuel:**

A very primitive drawing made in Screen-4. We labeled the value true or fake, URL and adding the database.



The problem both for this 3 section is Sentiment, Type, Topic Analysis labelling

So the app must give permit to do this manuel by the user till Phase-2 will finish or we can give permission to the user to select one of them from the database till Phase-2.

*(Proper nouns, starting words and subjects (we may change this into the metadata labelled in site for header.) Do you have any idea for this?*