There are many studies have been published on the topic of discriminate Arabic dialect, we can divide our reading in these studies to two main approaches:

1. **Researches that mainly concerns about identification between similar languages that are not Arabic.**

Many papers was published on this context, below is a short summary for each one of them.

* One of good readings in this area was published by AlinaMariaCiobanu, SergiuNisioi and LiviuP.Dinu (2016)[1], they worked on dataset of 20,000 sentence in each language or variety (18,000 training + 2,000 development), and 1000 for each language as a testing set.

- The goal for this paper was to identify the language of given sentence, languages are 13 here, and they divided to main groups, each group contain similar languages.

- They used several models, like one level logistic regression, but the best one was two level logistic regression, so that the first level predict in which group is the sentence and the second level predict the single dialect or language in that group. However, n-gram character was used as a feature extraction stage.

- Accuracy obtained in this system reached 89.9% in the best run which is impressive.

* Another method used for optimize such these systems done by Zampieri and Gebre (2015)[8], they work on the same problem of identifying 13 languages.

They used 3 approaches to develop the best system:

1. Logistic Regression with TF-IDF Weighting

2. SVM with TF-IDF Weighting

3. Likelihood Estimation

And the best of them was SVM with TF-IDF Weighting with maximum accuracy of 95.24%.

* Many other researches was published such as Maier’s paper (2014) [9] he made model on tweets for Spanish varieties, and he used character n-gram as a feature extraction method.

1. **Researches that concerns about identification between Arabic dialects.**

**We can this section to two main branches:**

* **Researches that come to solve same problem that we are working on.**
* One of the best and recent papers in this context is Arabic Dialect Identification in Speech Transcripts for (Shervin Malmasi, Marcos Zampieri)[2].

This paper won the first place in the DSL2016 competition, with accuracy of 51%.

They used 7 different feature extraction methods to get the feature vectors, the methods was 1-6 character n-gram and word uni-gram, SVM model was built from each of these feature extraction techniques.

Moreover, to predict any new instance, they use 3 different techniques:

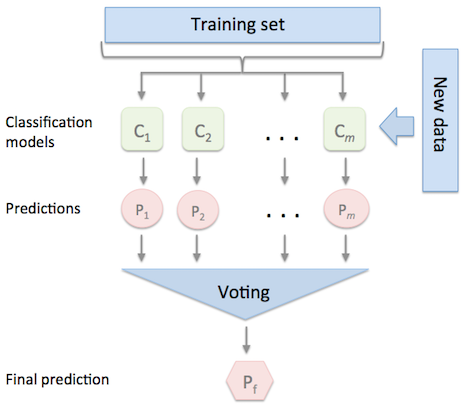


Figure 2.1: idea of voting between classifiers[3]

1. System #1: Pure voting (majority): in this method, each classifier will predict a single label, then the label with highest number of votes wins, other wise, the output will be selected randomly, this system made an accuracy of 49.16%.

2. System #2: Median probability ensemble: in this method, there is a probability assign to each label by each and every classifier, then we can find the median probability for each class, finally, label with largest median will wins, this system made an accuracy of 49.29%.

3. System #3: Mean probability ensemble: in this method, the same idea of last approach applied, the only difference is that mean probability deals with average probability instead of median probability, this system made an accuracy of 51.17%.

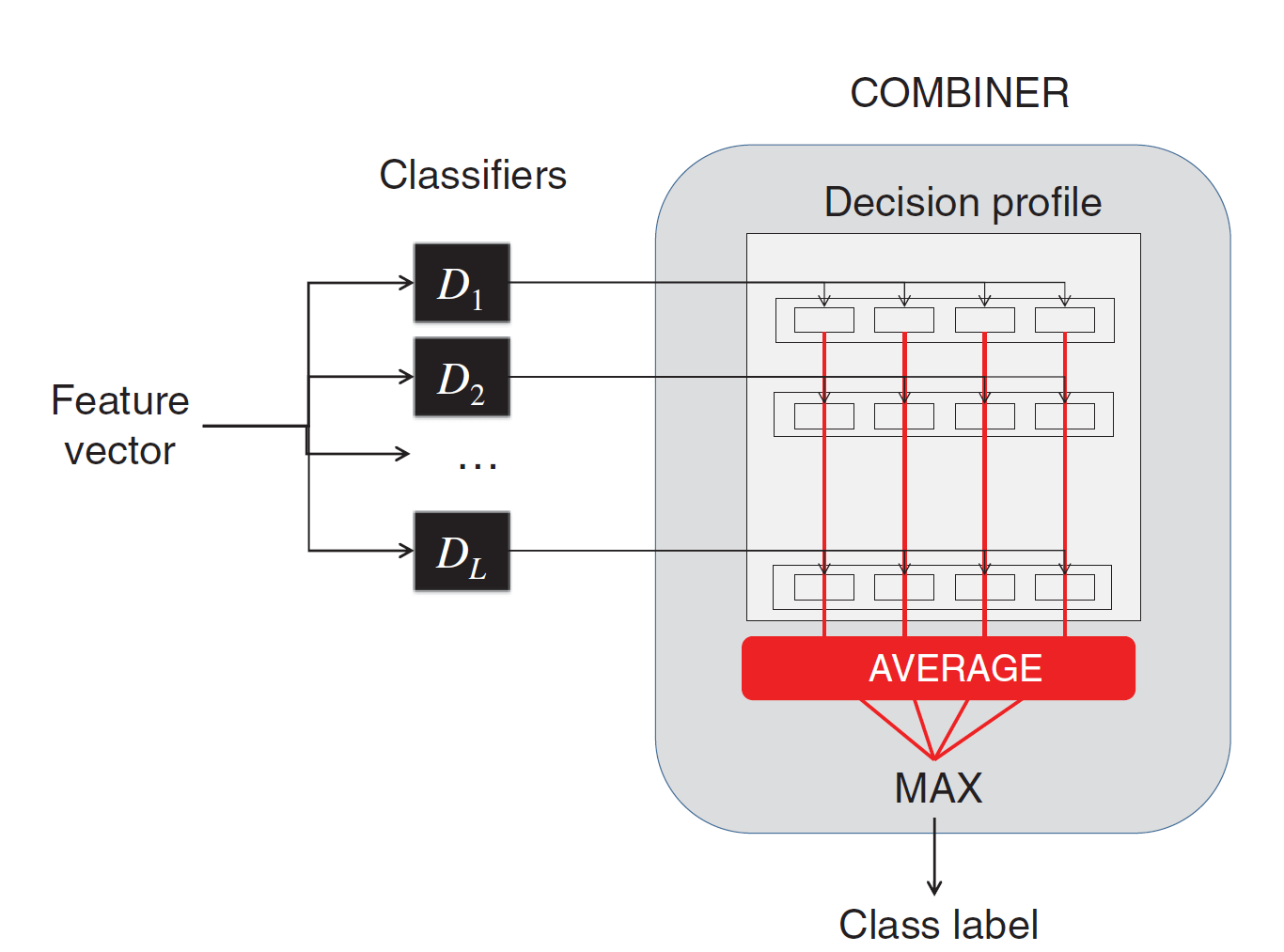


Figure2.2: An example of how mean probability ensemble works (for 4 label classification problem), each class take a probability from each classifier and then find the max from all averages.[4]

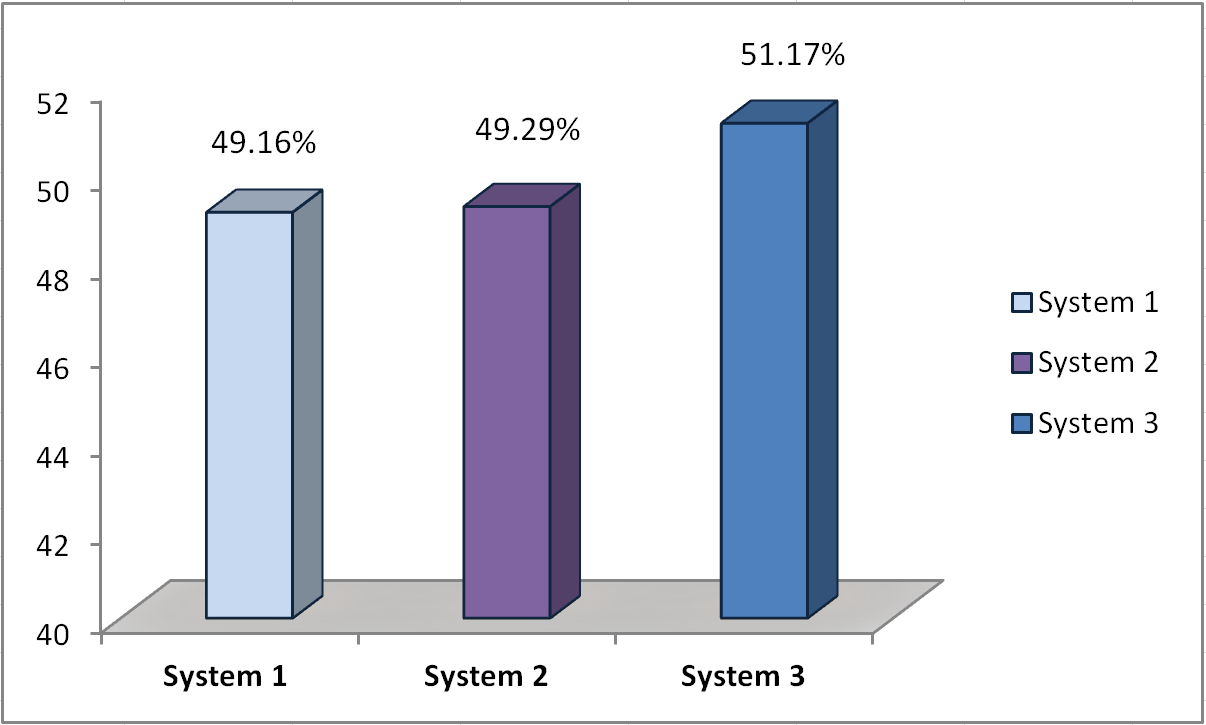


Figure 2.3:( Malmasi and Zampieri) Systems accuracies.

* Another rich paper in this context was for Mohamed Eldesouki(2014)[5] and his’s team in same competition, they got the third place with accuracy of 0.5136 and F1 weight of 0.5112.

They used several classification and features extraction techniques, the best was using (1-5) character n-gram with tf-idf as a features extraction method, and linear SVM as a classifier.

* **Researches that come so solve any Arabic dialects identification problem.**
* There are many papers studies Arab dialects identification, one of them was for Elfardy and Diab(2013)[6]. They built a system for identification between Egypt dialect and Standard Arabic language. Using Arabic **online-commentary** dataset they built a Naïve-Bayes model achieved accuracy of 85.5%.

However, **online-commentary** is a dataset collected from users commentaries on Egyptian news articles. See Fig2.4.

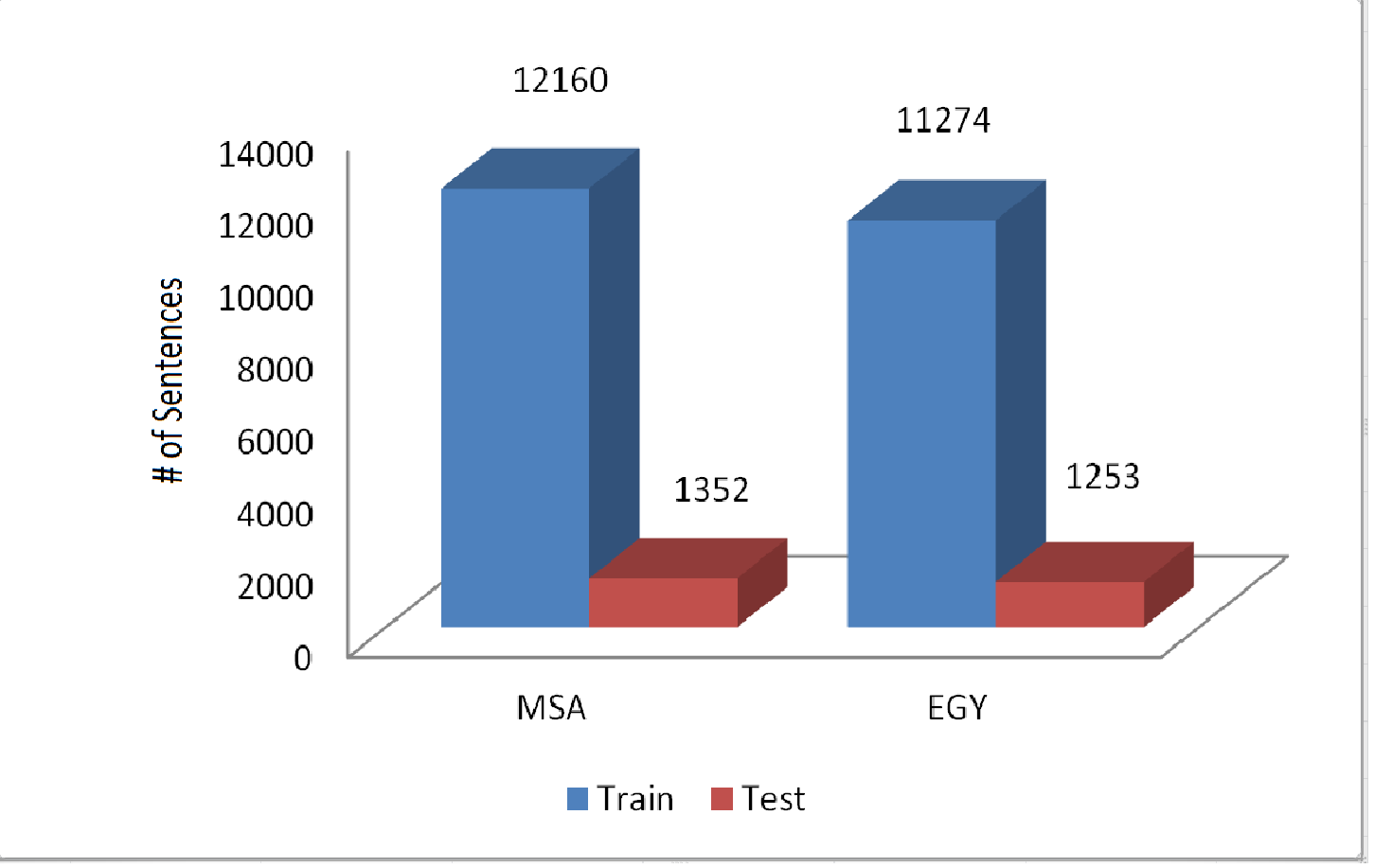


Fig2.4: train and test data used by Elfardy

* Another helpful reading in this discipline was by Darwish and Sajjad (2014), they use interesting technique to discriminate Arabic EGY dialect and MSA (standard Arabic).

The method built on the idea of using lexical, morphological, and phonological phenomena of dialects to identify that dialect, and they made some examples for each type of these features.(the Examples got from the Paper as It was).

* Dialectal words: that means EGY has some words that do not overlap with standard Arabic at all.

Example: “n$wf” (means: we see).

* Morphologicaldifferences: EGY make some Morphological pattern that in not part of MSA pattern.

Example: MSA: “ylEb” (he plays)→EG: “bylEb”.

* Letter substitution: It’s clear that dialects sometimes substitute or delete some letters from standard language.

Example: “v”→“t”. Ex. MSA: “kvyr” (a lot)→EG: “ktyr”.

* Syntactic differences: like a common use of masculine plural or singular noun forms instead dual and feminine plural.

Example: MSA: “metran” (two meter)→EG: “Atnyn mtr”.

* Taking these features in account, made them able to achieve 94.4% accuracy, they used dataset from International Workshop on Arabic Language Translation6 which provide approximately 150k sentence.
* There are many other researches like Malmasi(2015)[7], he built a system to discriminate between MSA and a group of Arabic dialect which are : {Egyptian, Jordanian, Palestinian, Syrian, and Tunisian} Parallel Multidialectal Corpus.

**[1:]** [**https://github.com/AhmadDarKhalil/project/blob/master/papers/VarDial330.pdf**](https://github.com/AhmadDarKhalil/project/blob/master/papers/VarDial330.pdf)

**[2]:** [**https://github.com/AhmadDarKhalil/project/blob/master/papers/Firtst.pdf**](https://github.com/AhmadDarKhalil/project/blob/master/papers/Firtst.pdf)

**[3]:** [**http://rasbt.github.io/mlxtend/user\_guide/classifier/EnsembleVoteClassifier**](http://rasbt.github.io/mlxtend/user_guide/classifier/EnsembleVoteClassifier)

**[4]:**[**https://github.com/AhmadDarKhalil/project/blob/master/papers/malmasi2015Votin g.pdf**](https://github.com/AhmadDarKhalil/project/blob/master/papers/malmasi2015Votin%20g.pdf)

**[5]:** [**https://github.com/AhmadDarKhalil/project/blob/master/papers/First%20Place.pdf**](https://github.com/AhmadDarKhalil/project/blob/master/papers/First%20Place.pdf)

**[6]:** [**https://github.com/AhmadDarKhalil/project/blob/master/papers/MSA\_Egy\_Heba.pdf**](https://github.com/AhmadDarKhalil/project/blob/master/papers/MSA_Egy_Heba.pdf)

**[7]:** Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015. Arabic Dialect Identification using a Parallel Multidialectal Corpus. In Proceedings of PACLING.

[8]: [**https://github.com/AhmadDarKhalil/project/blob/master/papers/W15-5411.pdf**](https://github.com/AhmadDarKhalil/project/blob/master/papers/W15-5411.pdf)

[9]: Wolfgang Maier and Carlos G´omez-Rodr´ıguez. 2014. Language Variety Identiﬁcation in Spanish Tweets. In Proceedings of the Workshop on Language Technology for Closely Related Languages and Language Variants, LT4CloseLang 2014, pages 25–35.