CIS 530: Project - Identifying Words to Anonymize in Spanish Medical Corpus
MILESTONE 3

Introduction -

We have built a baseline model simulating the kind of work implemented in the paper on 'Identifying Personal Health Information Using Support Vector Machines' (http://staffwww.dcs.shef.ac.uk/people/R.Gaizauskas/research/papers/amia06-deident.pdf). Since our dataset is in Spanish which is a low-resource language, we couldn't implement some features because of the scarcity of data for this particular locale. We intend to include more features based off of Spanish as an extension in Milestone 4. We have tried to gather as many token-level features as we could find for Spanish in this milestone.

Feature Set -

0. word
1. Word length
2. Whether word contains 1 digit
3. Whether word contains 2 digits
4. Whether word contains 3 digits
5. Whether word contains 5 digits
6. Whether word contains 6 digits
7. Whether word contains 7 digits
8. Whether word contains 9 digits
9. Whether there's an uppercase letter
10. Whether the word contains punctuations
11. Whether the word is Roman
12. Whether the word contains 'edad' or 'años'
13. Whether word is all uppercase
14. Whether word is all lowercase
15. Whether an apostrophe is present in the word
16. Whether a dash is present in the word
17. Whether the word contains 'fax'

For each word we look at the window [-1,0,1,2] and create feature vector including the features of these words in the context of the target word.

We have picked these particular features because:

- Digits:
    - Spanish phone numbers contain 9 digits and their format varies. It may come in 3 3 3, 3-3-3, 9, 2 3 2 2, 3 2 2 2, 3 6, 2 7
    - That is why we created a feature to capture all these different formats
- Upper/Lower cases To extract words that are part of names or addresses
- Edad/Anos to get values associated with age

- To distinguish between phone and fax numbers

Experiments with features:

We have a total of 401 training documents, 193 dev documents and 156 test documents and we tokenized the files into sentences who counts are mentioned as follows -

train_sents len= 8300 (Total (401 docs))
dev_sents len= 4048 (Total (193 docs))
test_sents len= 3231 (Total (156 docs))

Model used = LinearSVC

Experiment 1: features list: 0, 1
Experiment 2: features list: 0, 1, 2
Experiment 3: features list: 0, 1, 2, 3, 4, 5
Experiment 4: features list: 0, 1, 2, 3, 4, 5, 6, 7, 8
Experiment 5: features list: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
Experiment 6: features list: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17

Table 1. Table of all the experiments with features

	Train Set			**Dev Set**			**Test Set**		
Experiment	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
1	0.9784	0.9391	0.9583	0.7933	0.7165	0.753	0.7954	0.726	0.7591
2	0.9788	0.9396	0.9588	0.7938	0.7168	0.7533	0.7961	0.7271	0.76
3	0.9791	0.9398	0.9591	0.8178	0.7462	0.7804	0.8224	0.7544	0.7869
4	0.979	0.9396	0.9589	0.8544	0.7814	0.8163	0.8616	0.7914	0.825
5	0.9796	0.9403	0.9595	0.8865	0.8214	0.8527	0.8933	0.8313	0.8612
6	**0.9792**	**0.9406**	**0.9595**	**0.8868**	**0.8196**	**0.8519**	**0.8915**	**0.8335**	**0.8615**

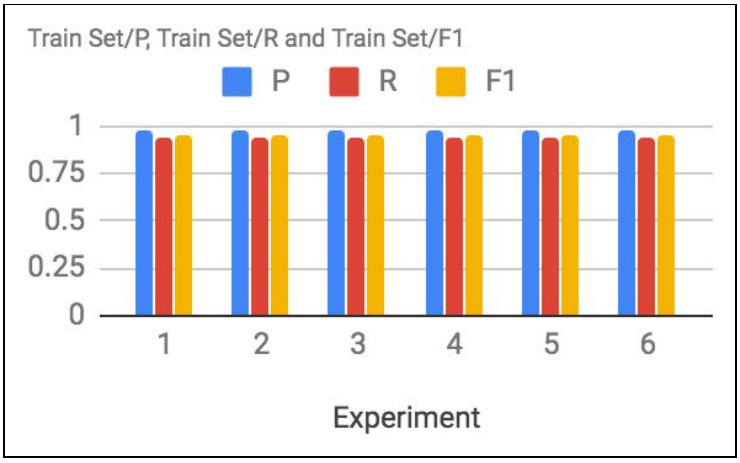

Figure 1.

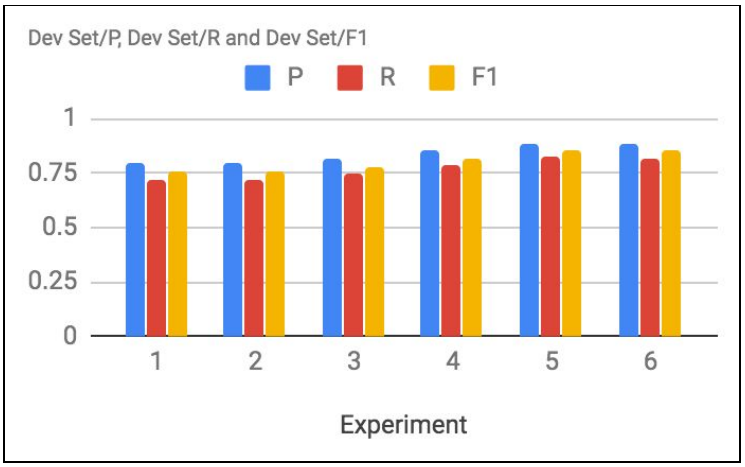

Figure 2.

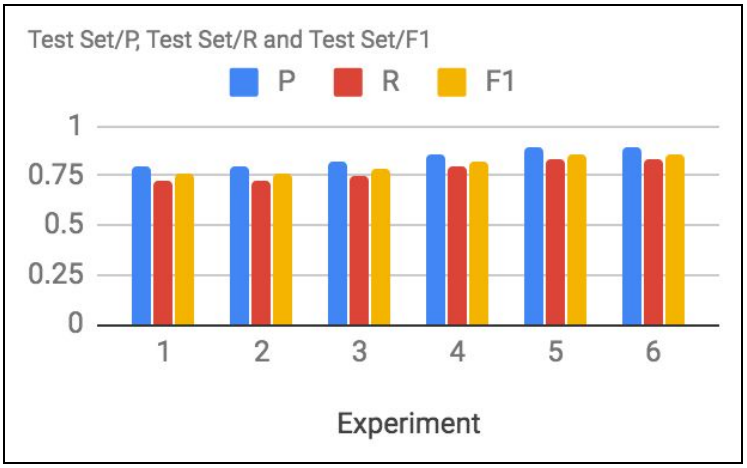

Figure 3.

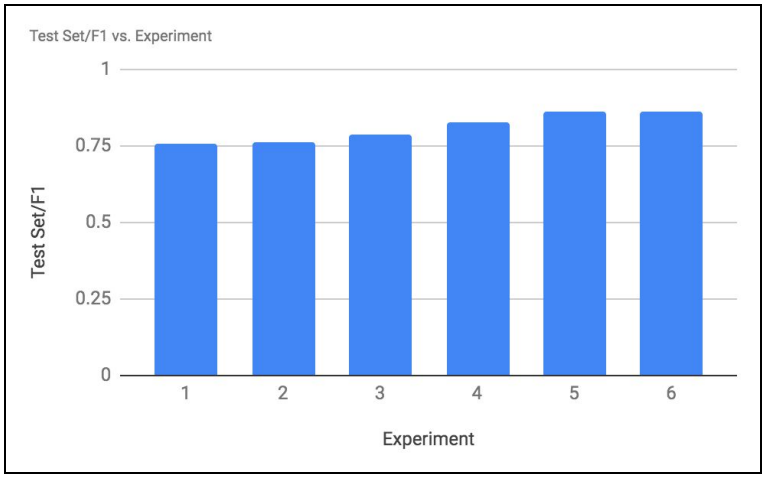

Figure 4.

Table 2. Evaluation metrics for individual categories

Train Data			
Category	**Precision**	**Recall**	**F1-score**
CALLE	94.50%	98.60%	96.51
CENTRO_SALUD	100.00%	100.00%	100
CORREO_ELECTRONICO	90.74%	99.40%	94.87
EDAD_SUJETO_ASISTENCIA	93.60%	97.65%	95.58
FAMILIARES_SUJETO_ASISTENCIA	92.22%	89.02%	90.59
FECHAS	99.80%	99.49%	99.64
HOSPITAL	88.54%	97.70%	92.9
ID_ASEGURAMIENTO	100.00%	100.00%	100
ID_CONTACTO_ASISTENCIAL	100.00%	98.08%	99.03
ID_SUJETO_ASISTENCIA	98.59%	98.82%	98.7
ID_TITULACION_PERSONAL_SANITARIO	99.72%	99.72%	99.72
INSTITUCION	93.94%	96.88%	95.38
NOMBRE_PERSONAL_SANITARIO	99.11%	99.49%	99.3
NOMBRE_SUJETO_ASISTENCIA	100.00%	99.63%	99.81
NUMERO_FAX	85.71%	85.71%	85.71
NUMERO_TELEFONO	95.35%	100.00%	97.62
OTROS_SUJETO_ASISTENCIA	75.00%	85.71%	80
PAIS	96.76%	99.63%	98.18

PROFESION	100.00%	100.00%	100
SEXO_SUJETO_ASISTENCIA	99.72%	99.86%	99.79
TERRITORIO	97.45%	99.57%	98.5

Table 3. Evaluation metrics for individual categories

Dev Data			
Category	**Precision**	**Recall**	**F1-score**
CALLE	35.77%	57.10%	43.99
CENTRO_SALUD	0.00%	0.00%	0
CORREO_ELECTRONICO	92.35%	94.94%	93.63
EDAD_SUJETO_ASISTENCIA	90.63%	92.27%	91.44
FAMILIARES_SUJETO_ASISTENCIA	64.71%	60.27%	62.41
FECHAS	94.11%	92.18%	93.13
HOSPITAL	39.39%	53.06%	45.22
ID_ASEGURAMIENTO	99.22%	95.49%	97.32
ID_CONTACTO_ASISTENCIAL	0.00%	0.00%	0
ID_EMPLEO_PERSONAL_SANITARIO	0.00%	0.00%	0
ID_SUJETO_ASISTENCIA	84.43%	84.83%	84.63
ID_TITULACION_PERSONAL_SANITARIO	96.47%	98.20%	97.33
INSTITUCION	19.05%	13.79%	16
NOMBRE_PERSONAL_SANITARIO	72.28%	85.56%	78.37
NOMBRE_SUJETO_ASISTENCIA	85.02%	95.35%	89.89
NUMERO_FAX	100.00%	50.00%	66.67
NUMERO_TELEFONO	70.83%	77.27%	73.91
OTROS_SUJETO_ASISTENCIA	100.00%	25.00%	40
PAIS	97.33%	96.23%	96.77
PROFESION	25.00%	25.00%	25
SEXO_SUJETO_ASISTENCIA	98.55%	98.83%	98.69
TERRITORIO	90.72%	83.24%	86.82

Table 4. Evaluation metrics for individual categories

Test Data			
Category	**Precision**	**Recall**	**F1-score**
CALLE	37.73%	60.82%	46.57
CENTRO_SALUD	20.00%	33.33%	25
CORREO_ELECTRONICO	88.08%	98.52%	93.01
EDAD_SUJETO_ASISTENCIA	88.13%	91.38%	89.73
FAMILIARES_SUJETO_ASISTENCIA	64.47%	56.98%	60.49
FECHAS	97.19%	98.45%	97.81
HOSPITAL	36.36%	48.89%	41.71
ID_ASEGURAMIENTO	98.43%	97.66%	98.04
ID_CONTACTO_ASISTENCIAL	0.00%	0.00%	0
ID_SUJETO_ASISTENCIA	82.86%	88.96%	85.8
ID_TITULACION_PERSONAL_SANITARIO	96.45%	99.27%	97.84
INSTITUCION	4.00%	2.78%	3.28
NOMBRE_PERSONAL_SANITARIO	74.44%	87.58%	80.48
NOMBRE_SUJETO_ASISTENCIA	85.51%	95.86%	90.39
NUMERO_FAX	0.00%	0.00%	0
NUMERO_TELEFONO	53.33%	66.67%	59.26
OTROS_SUJETO_ASISTENCIA	0.00%	0.00%	0
PAIS	97.24%	95.91%	96.57
PROFESION	25.00%	25.00%	25
SEXO_SUJETO_ASISTENCIA	97.25%	100.00%	98.61
TERRITORIO	90.02%	84.50%	87.17

Comparison with Random Baseline:

Model	**Train Set**			**Dev Set**			**Test Set**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Random Baseline	0.634	0.4382	0.5182	0.6292	0.4254	0.5076	0.628	0.4238	0.5061
M3 Baseline	0.9792	0.9406	0.9595	0.8868	0.8196	0.8519	0.8915	0.8335	0.8615

Link for the final presentation :

https://docs.google.com/presentation/d/1fWq5xAv3SrCDAZkWzhFfLejaNQbNutaWzkpx2fmtTmY/edit#slide=id.g58186f86f5_1_3