An Exploratory Study on the Impact of Temporal Features on the Classification and Clustering of Future-Related Web Documents

Ricardo Campos [2, 3, 4], Gaël Dias [1, 2], Alípio Jorge [4, 5]

[1] DLU/GREYC, Univeristy of Caen Basse-Normandie, Caen, France

[2] Centre of Human Language Technology and Bioinformatics, University of Beira Interior, Covilhã, Portugal

[3] Tomar Polytechnic Institute, Tomar, Portugal

[4] LIAAD-INESC Porto L.A , Porto, Portugal

[5] Faculty of Sciences, University of Porto, Porto, Portugal

EPIA 2011 – 15th Portuguese Conference on Artificial Intelligence, Lisbon - Portugal, October 10 - 13, 2011

[www.ipt.pt] [www.liaad.up.pt] hultig.di.ubi.pt]

Welcome to the World Wide Web

European Economic Forecast further puts emphasis on predictions…to be rather high in **2011** and could further accelerate during **2012**.

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Queries that Combine Text / Time

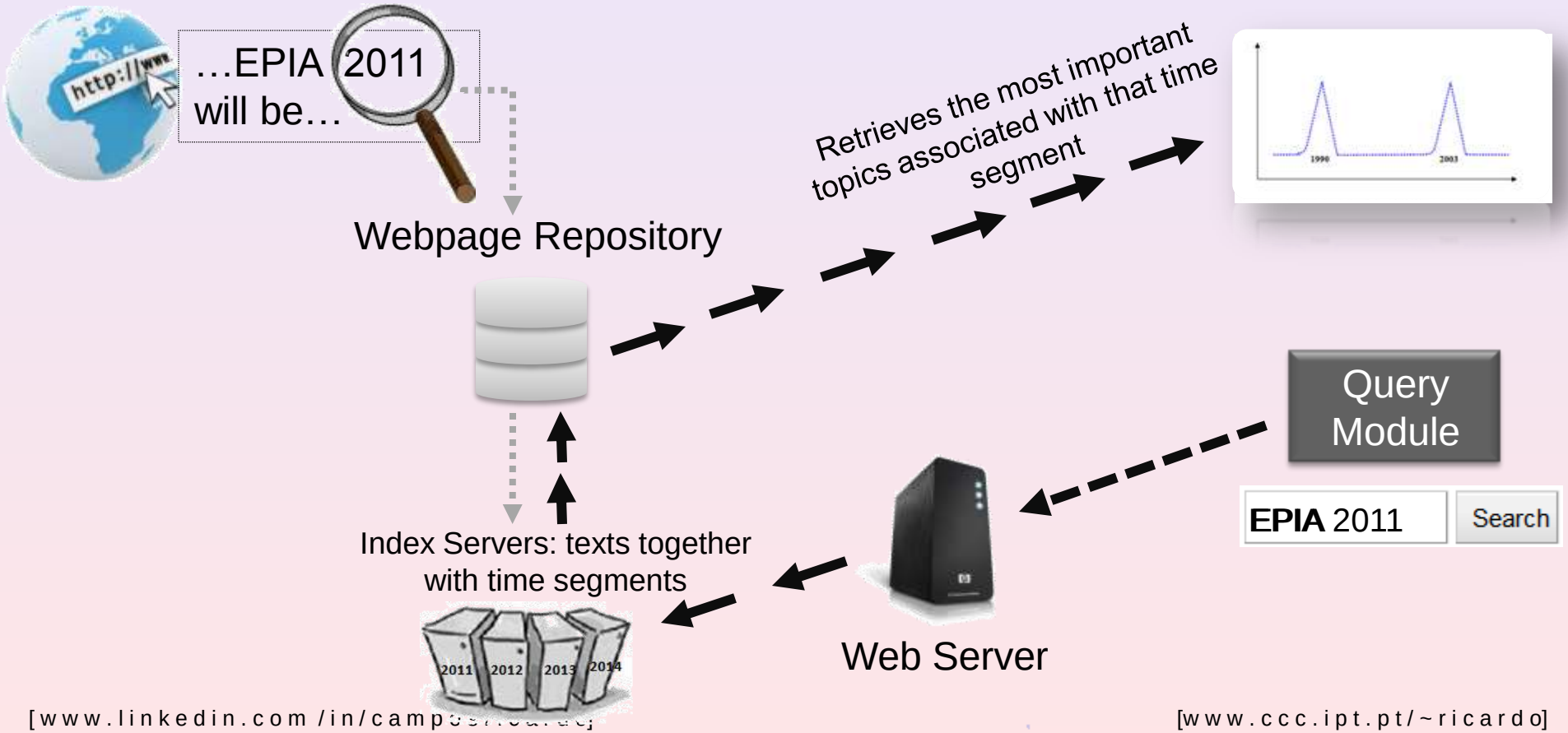
Dacia plans the release of no less than 8 new models and facelifts by **2015**.

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Architecture of a Future Retrieval System

Overall a FR system should be composed of:



[www.linkedin.com/in/campos...]

[www.ccc.ipt.pt/~ricardo]

Searching for Future Temporal References

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Metadata-Based Approach

Jun 16, 2009 – The city of São Paulo shall have to make use of the Credicard Hall as the venue for the **2011 Miss Universe**. **Today** was also announced that **Miss Morumbi** show is going to be on **July 27, 2009.**
From Miss Universe.Com

Content-Based Approach

Tour de France 2011 - Grand start **2012**
www.letour.fr/2011/TDF/.../us/grand_depart_**2012**.html
It is with enthusiasm and undisguised pride that once again the Province of Liège will host the "Grand Départ" of the **2012 Tour de France** - after the **...**

However, in contrast to web news articles, which are mainly of

web documents, especially those from social networks, **suffer** however **from the problem** of containing a **large number of**

- Informative;
- Schedule;
- Rumor;

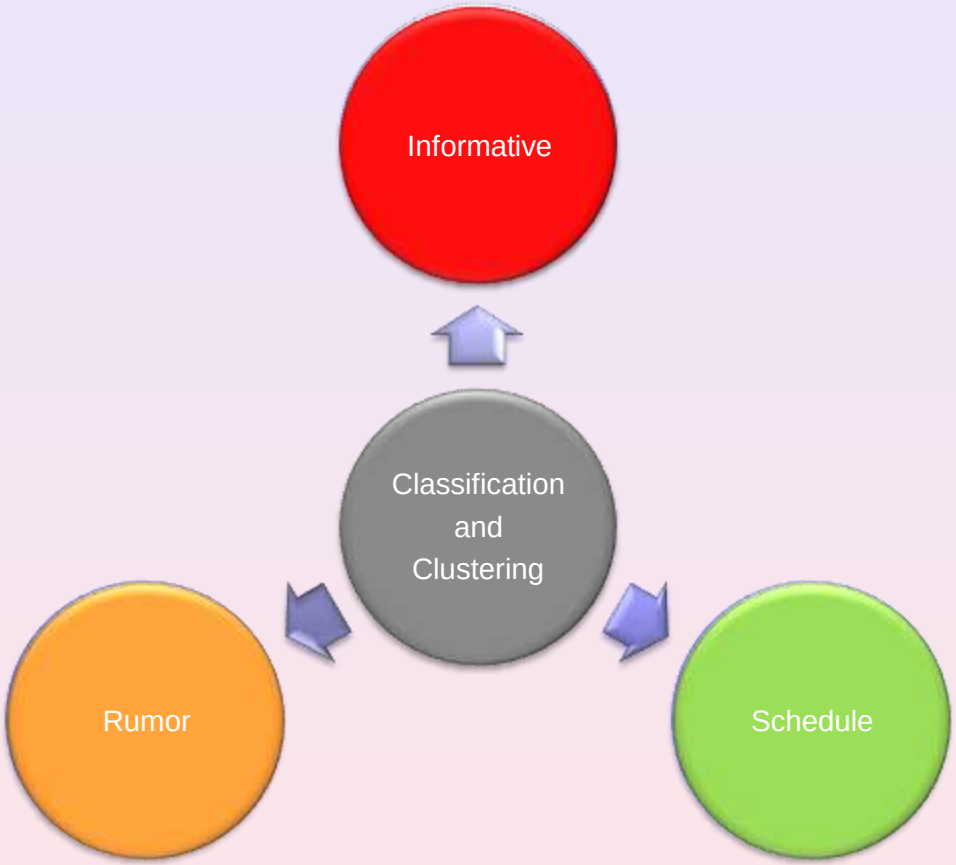
Hairstyles - Haircuts for **2012**, Hair Styles and Makeup tips

The **2014** FIFA World Cup, is an international football tournament that is scheduled to take place in June and July **2014** in Brazil.

Many predictions of the end of the world in **2012** have seriously scared some people.

[www.linkedin.com /in/camposricardo]
[www.ccc.ipt.pt/~ricardo]

Understand the Impact of Temporal Features



Following a Content-Based approach

- We are particularly interested in considering a specific type of query, the so-called **implicit temporal query:**

Dacia Duster | Search

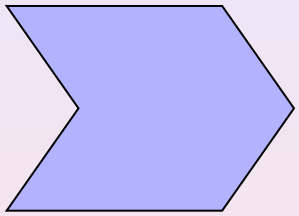
- We are mainly focused on **year dates** based on the analysis of **web content** within web snippets:

Alice in Wonderland (2010 film) - Wikipedia, the free encyclopedia
Alice in Wonderland is a 2010 American computer-animated/live action fantasy
en.wikipedia.org/wiki/**Alice_in_Wonderland**_(2010_film)

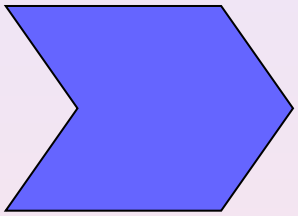
[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

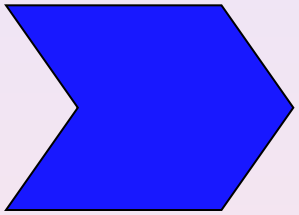
Summary



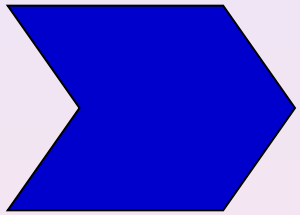
Future Temporal Nature of Web Documetns



Classification of Future-Related Texts



Clustering of Future-Related Texts



Conclusions

450 Implicit Temporal Queries. 62.842 Web Snippets

Insights for Search

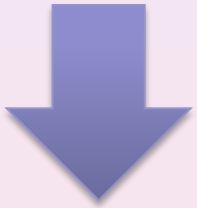
20 queries

* 27 categories

540 queries

January 2010 – October 2010

Oil Spill;
BP Oil Spill;
Dacia;



Alice in Wonderland (2010 film) - Wikipedia, the free encyclopedia
Alice in Wonderland is a 2010 American computer-animated/live action fantasy
en.wikipedia.org/wiki/Alice_in_Wonderland_(2010_film)

62.842 Web Snippets

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Automatic Date Identification

	Dates		Future Dates			Near Future Dates		
				A	R		A	R
Snippets	5777	9.2%	508	0.8%	8.7%	419	0.6%	82.4%
Titles	2058	3.2%	419	0.6%	20.3%	373	0.5%	88.7%
URLs	3512	5.5%	195	0.3%	5.5%	167	0.2%	85.6%

- *FIFA plans to establish a base for the 2018-2022 World Cup leading; by **2015**;*
- *Japan plans to establish a base for the 2018-2022 World Cup leading; by **2015**;*
- The **occurrence of dates** is largely predominant in **2011**, but **consistent until 2013**;
- Thereafter, there are some quite **small peaks** in **2014** and **2022** that mostly relate to the Football World Cup.

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Occurrence of Future Dates

	Dates		Future Dates			Near Future Dates		
				A	R		A	R
Snippets	5777	9.2%	508	0.8%	8.7%	419	0.6%	82.4%
Titles	2058	3.2%	419	0.6%	20.3%	373	0.5%	88.7%
URLs	3512	5.5%	195	0.3%	5.5%	167	0.2%	85.6%

- Overall, the **occurrence of future dates** is very common in items retrieved **in response to queries** belonging to the categories of




Future Dates in Snippets

	Dates		Future Dates			Near Future Dates		
				A	R		A	R
Snippets	5777	9.2%	508	0.8%	8.7%	419	0.6%	82.4%
Titles	2058	3.2%	419	0.6%	20.3%	373	0.5%	88.7%
URLs	3512	5.5%	195	0.3%	5.5%	167	0.2%	85.6%

- Still, when talking in relative values, we can note that future dates occur in **8.79%** of the cases:

  *Honda is planning a major jump in hybrid sales in Japan in **2011;***

  *Avatar 2? in **2013**?*

  *IDC predicts sales of mobile apps will be a $35 billion industry by **2014;***

  *Wycliffe's mission is to see a Bible translation in every language by **2025***

[www.linkedin.com/in/camposricardo] [www.ccc.ipt.pt/~ricardo]

Future Dates in Titles

	Dates		Future Dates			Near Future Dates		
				A	R		A	R
Snippets	5777	9.2%	508	0.8%	8.7%	419	0.6%	82.4%
Titles	2058	3.2%	419	0.6%	20.3%	373	0.5%	88.7%
URLs	3512	5.5%	195	0.3%	5.5%	167	0.2%	85.6%

- Unlike in snippets, future dates are **very common in titles**. From a total number of 2.058 items tagged with dates, **20.3%** have a future temporal intent;
- and that about **90%** of the future dates are related to the **near future**

__2011__ will be best year to buy a home, says BSA;

Experts bet on India growth story in __2011__;

Tour de France organizers unveil climb-heavy __2011__ route;

Future Dates in URLs

	Dates		Future Dates			Near Future Dates		
				A	R		A	R
Snippets	5777	9.2%	508	0.8%	8.7%	419	0.6%	82.4%
Titles	2058	3.2%	419	0.6%	20.3%	373	0.5%	88.7%
URLs	3512	5.5%	195	0.3%	5.5%	167	0.2%	85.6%

- In the opposite side, the occurrence of future dates in URLs is scarce: **only 5.5%** of the links have a **future temporal nature**;
- but they are extremely descriptive:
  *http://www.grist.org/article/2010-11-15-fords-first-electric-car-to-be-sold-in-20-cities-in-**2011***

  *http://msn.foxsports.com/foxsoccer/worldcup/story/world-cup-bid-usa-loses-**2022**-world-cup-bid-to-qatar*

[www.linkedin.com/in/camposricardo] [www.ccc.ipt.pt/~ricardo]

Classification of Texts according to Genre

#Item	#Items	Informative		Schedule		Rumors	
Snippets	508	255	50.2%	136	26.8%	117	23.0%
Titles	419	248	59.2%	85	20.3%	86	20.5%
URLs	195	101	51.8%	38	19.5%	56	28.7%

- At about 70% - 75% of the texts have either an **informative** nature or concern a **scheduled event** which has a very high probability of taking place:

  *Latest Hairstyles **2011**. Tickets for Lady Gaga **2011** Tour.*

- The remaining relate to **rumor texts**, which lack confirmation in the future:

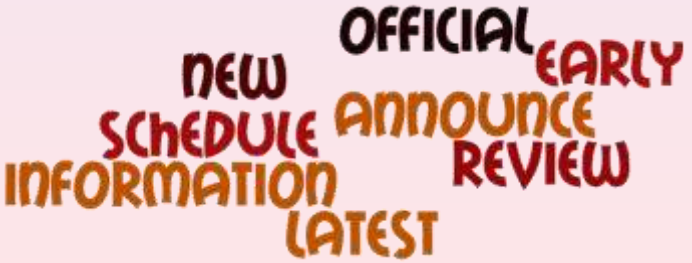
  *WebOS tablet will arrive in March **2011**. Details are not officially.*

[www.linkedin.com/in/camposricardo] [www.ccc.ipt.pt/~ricardo]

Classification of Texts according to Genre for Near/Distant Future Dates

Item	Near Future			Distant Future		
	Schedule	Informative	Rumor	Schedule	Informative	Rumor
Web Snippet	25.7%	55.8%	18.3%	31.4%	23.6%	44.9%
Title	15.0%	65.4%	19.5%	63.0%	8.7%	28.2%
URL	13.7%	56.8%	29.3%	53.5%	21.4%	25.0%

- **Informative texts** mostly occur with **near future dates**;
- Information on **product releases** (e.g., *Dacia Duster*, *Microsoft*) and upcoming **scheduled events** (e.g. *Auto Show*):



[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Classification of Texts according to Genre for Near/Distant Future Dates

Item	Near Future			Distant Future		
	Schedule	Informative	Rumor	Schedule	Informative	Rumor
Web Snippet	25.7%	55.8%	18.3%	31.4%	23.6%	44.9%
Title	15.0%	65.4%	19.5%	63.0%	8.7%	28.2%
URL	13.7%	56.8%	29.3%	53.5%	21.4%	25.0%

- **As we move forward** in the calendar, it is more common for texts to be **related to events** planned in advance and to also be of a **rumor nature**;

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Classification of Texts according to Genre for Near/Distant Future Dates

Item	Near Future			Distant Future		
	Schedule	Informative	Rumor	Schedule	Informative	Rumor
Web Snippet	25.7%	55.8%	18.3%	31.4%	23.6%	44.9%
Title	15.0%	65.4%	19.5%	63.0%	8.7%	28.2%
URL	13.7%	56.8%	29.3%	53.5%	21.4%	25.0%

- It is also important to note that **Future dates** are mostly **year related** and fewer are related to months or days. **Exceptions only occur with scheduled events**: *Tour de France: from Saturday July 2nd to Sunday July 24th 2011, the 98th*

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Our Goal

- We aim to understand whether **temporal features influence the classification and clustering of future-related texts** according to their nature: informative, scheduled or rumor.

- We need to have in mind that our **main objective is not to reach high accuracy results** but instead **to understand the impact of temporal features over different learning paradigms.**

- To understand whether the category of future-related documents can be discovered:

  by using only **specific linguistic features**;

  or if it can be improved **by including temporal features;**

Text Snippets

Experiments are based on two collections:

Focus Time

Near Future

Far Future

Text Genre

Informative

Schedule

Rumor

Focus Time

Near Future

Far Future

Text Genre

Informative

Schedule

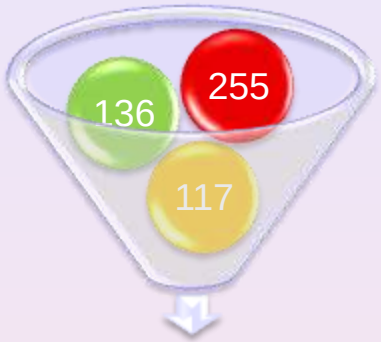
Rumor

[www.linkedin.com/in/camposricardo]

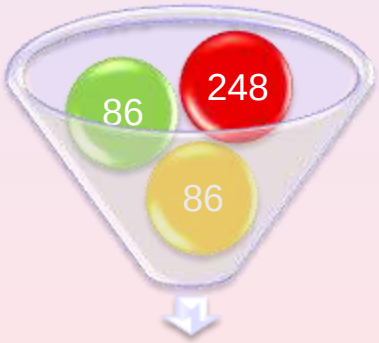
[www.ccc.ipt.pt/~ricardo]

Text Titles

The construction of balanced datasets resulted in the selection of:



117 + 117 + 117 = 351 Snippets



86 + 86 + 86 = 258 Titles

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Construction of 4 Balanced Datasets

For each one (text snippet dataset, and text title dataset) we built, respectively, four datasets:

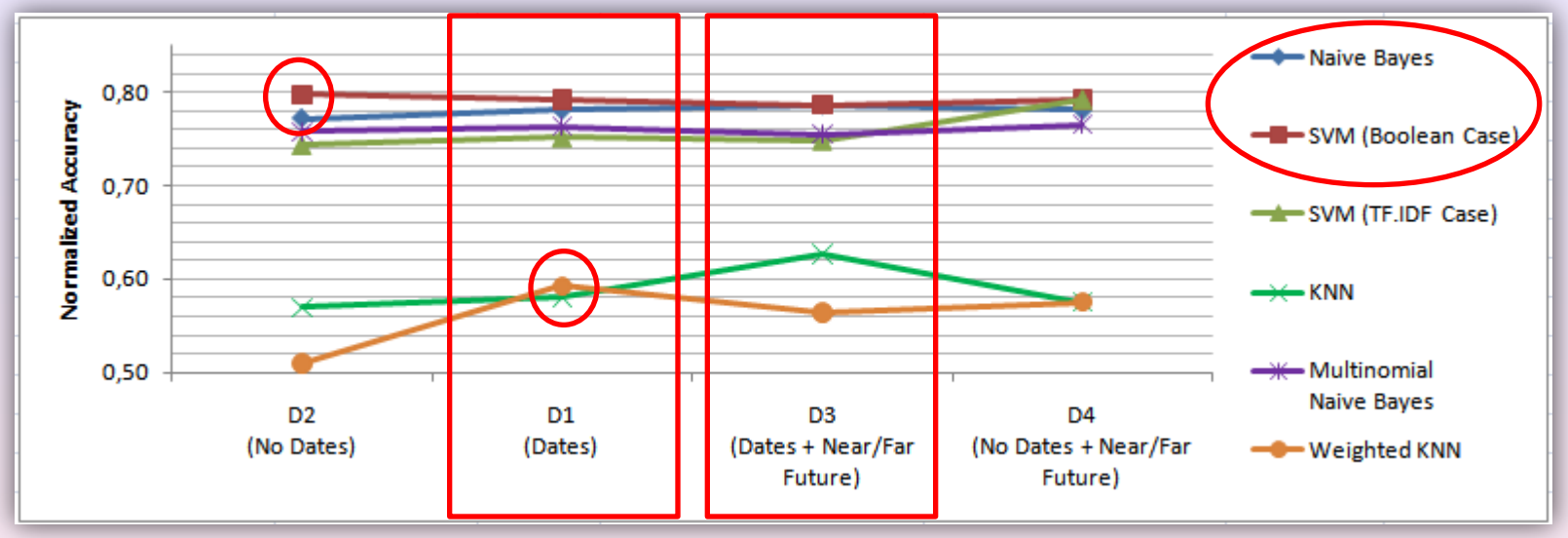
Dataset	Snippet / Title		Focus Time	Text Genre
	Text	Year Dates		
D1	x	x		x
D2	x			x
D3	x	x	x	x
D4	x		x	x

We used 5 Different Algorithms

Experiments were run on the basis of a **5-fold cross-validation** for **boolean** and **tf.idf unigram features** for **five different classifiers**:

- Naïve Bayes (boolean);
- Multinomial Naïve Bayes (tf.idf);
- K-NN (boolean). K = 10;
- Weighted K-NN (tf.idf). K = 10 and distance = (1 / weight);
- Multi-Class SVM (boolean and TF.IDF).

Text Snippets: Overall Analysis of Global Accuracy



All the algorithms, with the **exception of SVM (boolean) show improved results** with the simple **use of explicit year dates**

Weighted K-NN shows the highest difference

However, both **Naïve Bayes and SVM (boolean) largely outperform the Weighted K-NN** in terms of accuracy)

Best overall results are obtained for **SVM without any mention of time features**

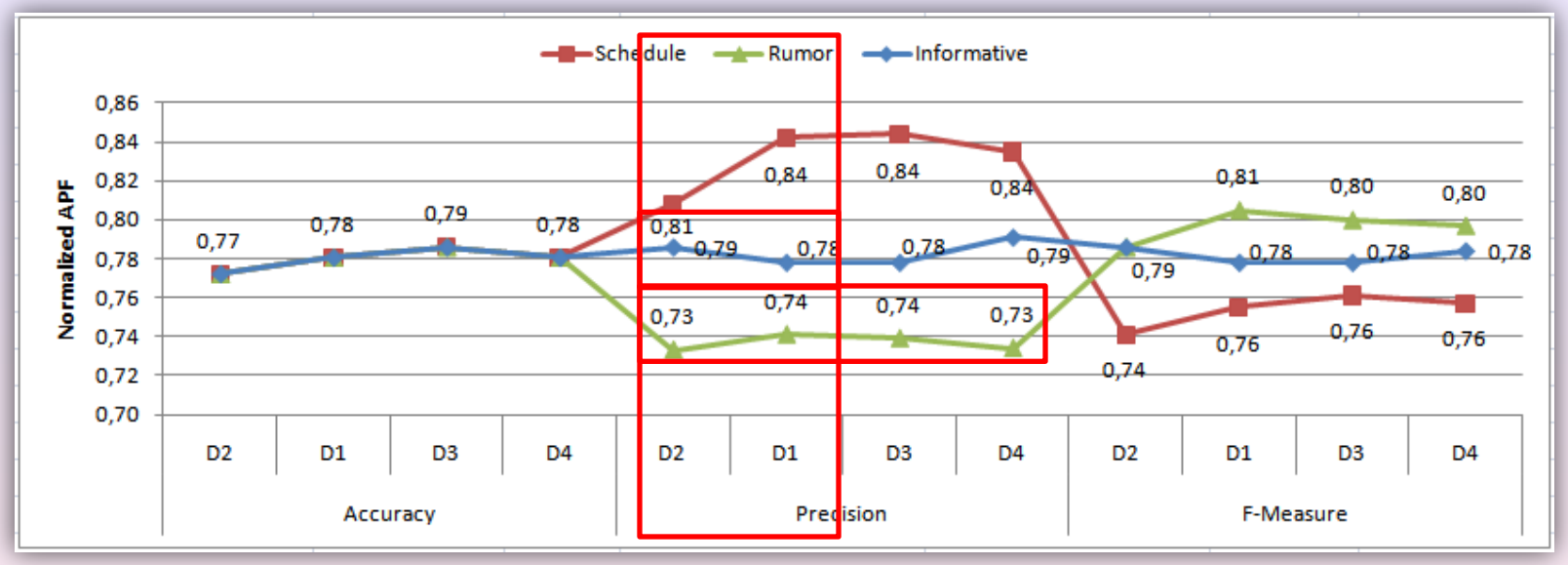
It seems that the **language used in each text genre is enough to classify future information**

Moreover, the **dates do not have a great impact if combined with near/distant future knowledge**

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Text Snippets: Specific Case of Naïve Bayes

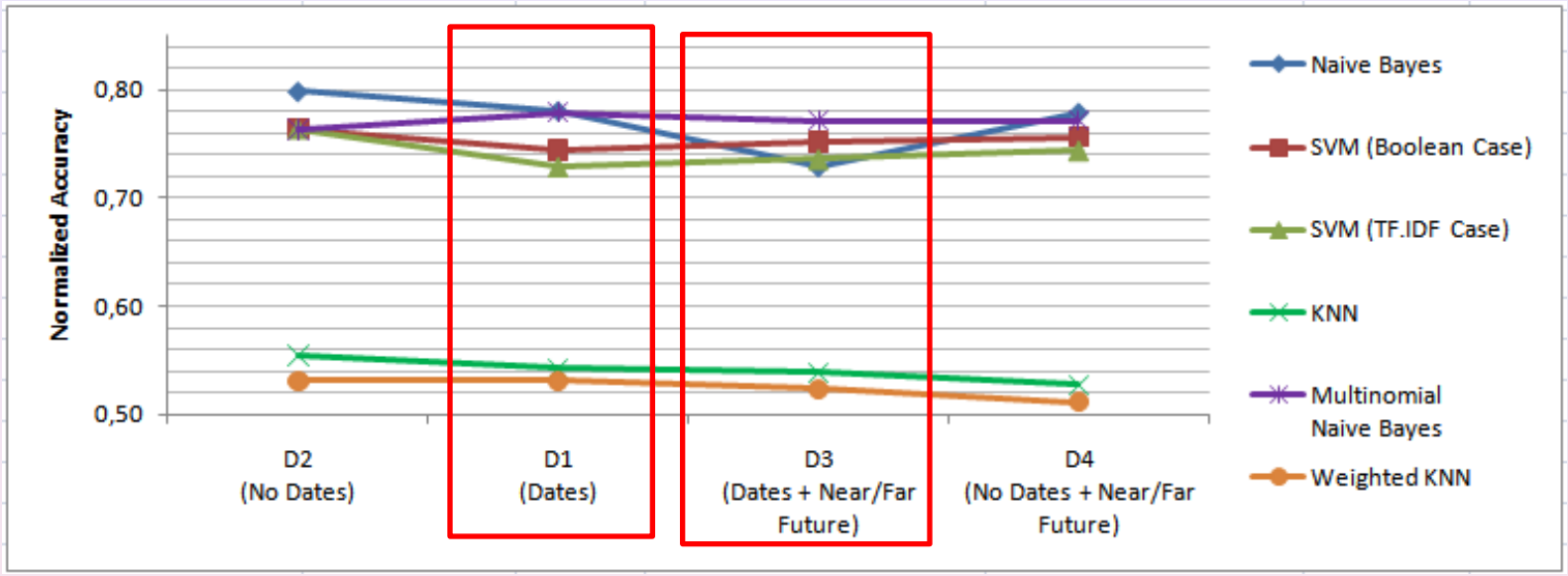


The introduction of **temporal features has an overall positive impact on precision in the classification of scheduled texts**

However, the classification of informative texts is more accurate without dates

is uncertain in the case of rumor texts

Text Titles: Overall Analysis of Global Accuracy



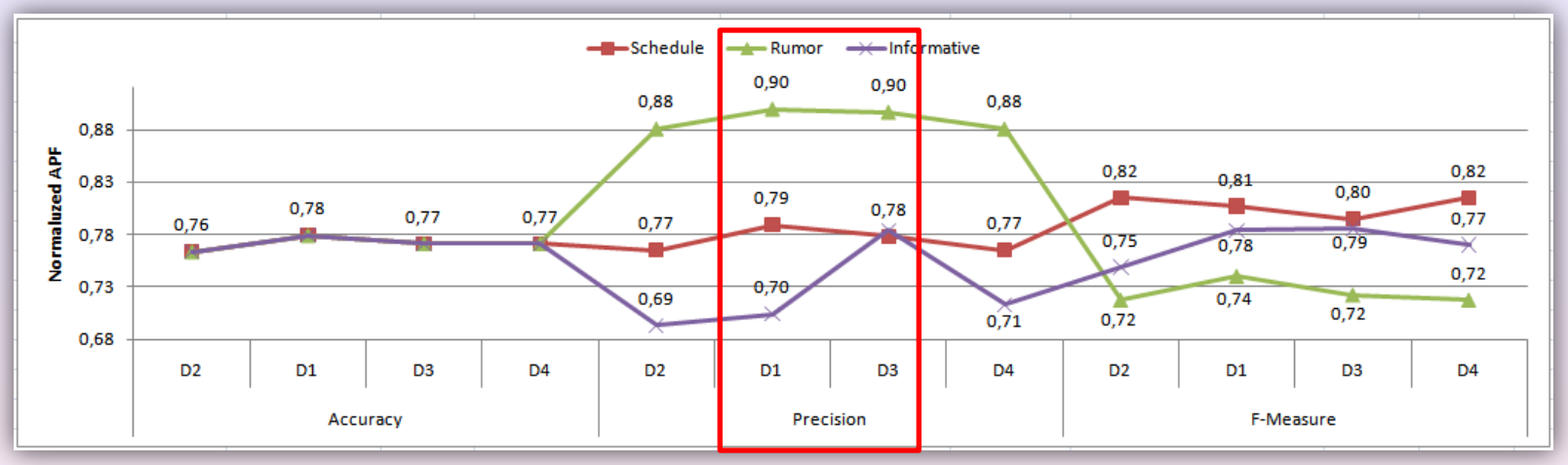
Overall, it is clear that **most of the algorithms perform worst in terms of accuracy with the introduction of temporal features**

This **means that time characteristics do not have a great impact on the classification task**

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Text Titles: Specific Case of Multinomial Naïve Bayes



Scheduled texts benefit with the introduction of temporal features

This is not as clear in the case of informative texts

We have detected that precision in rumor texts is very high

This just occur in the specific case of Multinomial Naïve Bayes. In all the others **time features do not have an overall impact on the classification task**

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

K-Means

To complete our study, we also proposed a set of experiments based on the well-known:

- K-Means;

The idea is to automatically retrieve **three different clusters** (informative, scheduled and rumors) based on the same representations of web snippets, the D1, D2, D3 and D4;

As in the classification case, experiments for the boolean and tf.idf cases, and for snippets and text titles are shown:

Text Snippets: Boolean and TF.IDF - Accuracy Results

Case	D2	D1 Year Dates	D3 Year Dates Focus Time	D4 Focus Time
Boolean	43.59%	43.59%	45.02%	41.88%
TF.IDF	35.90%	39.04%	40.74%	51.00%

For the boolean case the introduction of explicit year dates only improves the results when combined with the focus time

For the TF.IDF case the use of year dates improves the results

However, best results occur without year dates (considering only focus time information)

Text Titles: Boolean and TF.IDF - Accuracy Results

Case	D2	D1 Year Dates	D3 Year Dates Focus Time	D4 Focus Time
Boolean	42.25%	39.54%	39.54%	42.25%
TF.IDF	41.87%	41.87%	53.49%	41.87%

Best results occur for D3 in the tf.idf representation, with nearly a 13% impact when compared to D4

Rumor texts reach an impressive value of almost 85% in terms of precision

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Classification Task

Depending **on the representation of the snippet** and **on the algorithm**, the **temporal issue may or not have any influence**;

For the snippet **classification task**, the **SVM gives the overall best results** without **any temporal information** with **79.22% accuracy for the boolean case**;

Although the same **Multi-class SVM** shows improved results for the **tf.idf case** when the **near/far future date is introduced**, reaching **79.20% accuracy**;

Moreover, the **probabilistic learning** and the **lazy learning** families always evidence **best results** when any time feature is used, to the exception of the Multinomial Naive Bayes for D3;

Classification Task

This is the opposite of what happens with the **classification of text titles**, where most of the algorithms perform **better without temporal features**;

However in detail, we can also conclude that in general, **the introduction of temporal features** has an overall **positive impact** on the **classification of scheduled texts**, both in snippets as well as in text titles;

Interestingly we can also note that the **detection of rumor texts benefits from the introduction of temporal features**, particularly in the probabilistic algorithms;

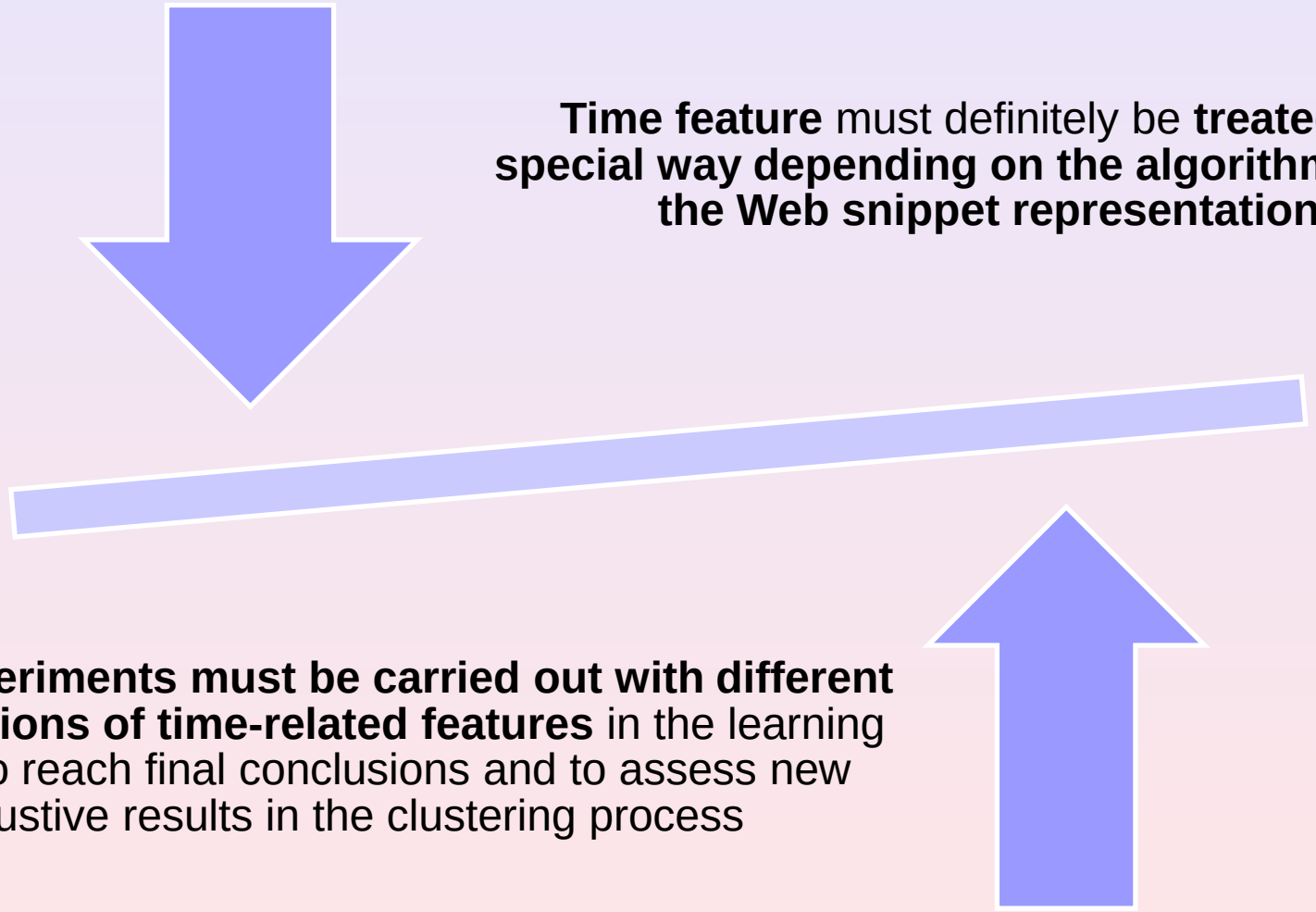
[www.linkedin.com/in/camposricardo] [www.ccc.ipt.pt/~ricardo]

Clustering Task

For the **clustering task**, the **impact of temporal features** is more apparent in **D1 for snippets** and in **D3 for text titles;**

Moreover, the identification of **schedule texts** is particularly easy in **text snippets**, while **rumor texts** are easily identified in text **titles**;

Further Experiments Must Be Carried Out



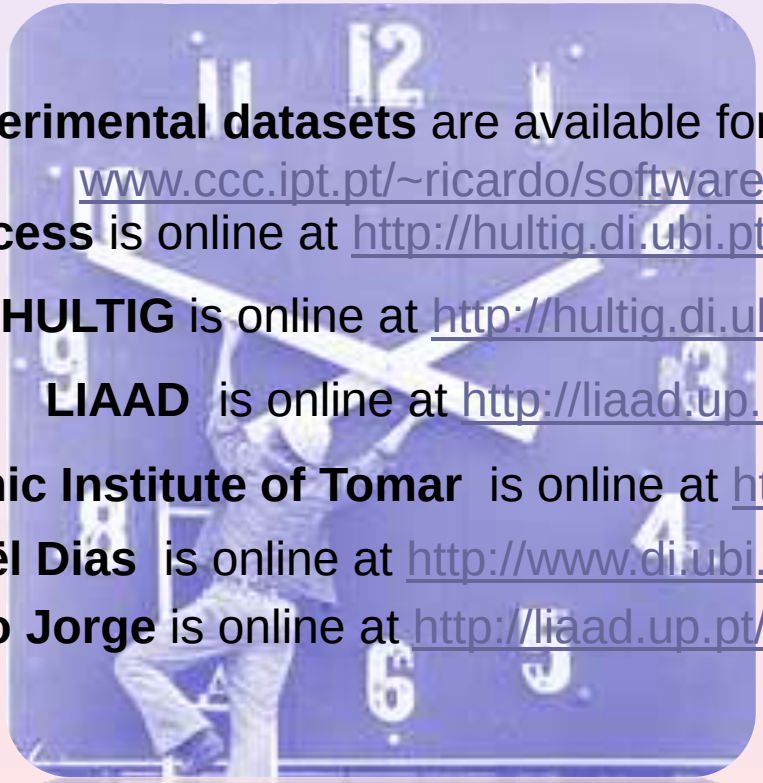
Time feature must definitely be **treated in a special way depending on the algorithm and on the Web snippet representation**;

Further experiments must be carried out with different representations of time-related features in the learning process, to reach final conclusions and to assess new exhaustive results in the clustering process

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]

Thanks for your attention!



Both **experimental datasets** are available for download at www.ccc.ipt.pt/~ricardo/software

VipAccess is online at http://hultig.di.ubi.pt/vipaccess

HULTIG is online at http://hultig.di.ubi.pt

LIAAD is online at http://liaad.up.pt

Polytechnic Institute of Tomar is online at http://www.ipt.pt

Gaël Dias is online at http://www.di.ubi.pt/~ddg

Alípio Jorge is online at http://liaad.up.pt/~amjorge

[www.linkedin.com/in/camposricardo]

[www.ccc.ipt.pt/~ricardo]