

A Supervised Term-Weighting Method and its Application

...

M. Maisonnave, F. Delbianco, F. Tohmé y A. Maguitman
47 JAIIO - 3 de Septiembre de 2018
Universidad Nacional del Sur

Motivations for Term Weighting

- Improve Information Retrieval Systems
- Text Representation for Classification

Term Importance is typically taken as a fixed value independent of the task at hand.

Motivations for Context-based Term Weighting

- Query formulation
- Term Relevance Scoring
- Variable Selection

Previous Work

Salton and Buckley (1988) claimed that at least **three main factors** are required in any term weighting scheme.

- **Local factor:** frequent terms are semantically close to the content of the document.
 - helps to improve recall.
- **Global Factor:** associated with each term, represents how frequent the term is in the document collection.
 - helps to improve precision.
- **Normalization Factor:** to penalize large documents.

Previous Work

Salton and Buckley (1988) claimed that at least **three main factors** are required in any term weighting scheme.

- **Local factor:** frequent terms are semantically close to the content of the document.
 - helps to improve recall.
- **Global Factor:** associated with each term, represents how frequent the term is in the document collection.
 - helps to improve precision.
- **Normalization Factor:** to penalize large documents.

Binary, TF,
ITF, ...

Previous Work

Salton and Buckley (1988) claimed that at least **three main factors** are required in any term weighting scheme.

- **Local factor:** frequent terms are semantically close to the content of the document.
 - helps to improve recall.
- **Global Factor:** associated with each term, represents how frequent the term is in the document collection.
 - helps to improve precision.
- **Normalization Factor:** to penalize large documents.

Binary, TF,
ITF, ...

Unsupervised: IDF, WIDF, ...

Previous Work

Salton and Buckley (1988) claimed that at least **three main factors** are required in any term weighting scheme.

- **Local factor:** frequent terms are semantically close to the content of the document.
 - helps to improve recall.
- **Global Factor:** associated with each term, represents how frequent the term is in the document collection.
 - helps to improve precision.
- **Normalization Factor:** to penalize large documents.

Binary, TF,
ITF, ...

Unsupervised: IDF, WIDF, ...
Supervised: ICF, MI, OR,
GSS, ...

Previous Work

Salton and Buckley (1988) claimed that at least **three main factors** are required in any term weighting scheme.

- **Local factor:** frequent terms are semantically close to the content of the document.
 - helps to improve recall.
- **Global Factor:** associated with each term, represents how frequent the term is in the document collection.
 - helps to improve precision.
- **Normalization Factor:** to penalize large documents.

Binary, TF,
ITF, ...

Unsupervised: IDF, WIDF, ...
Supervised: ICF, MI, OR,
GSS, ...

$$\mathbf{v}_{normalized} = \frac{1}{\sqrt{\sum_{i=1}^n v_i^2}} \times \mathbf{v}$$

Proposed Technique

DESCR

The **descriptive relevance** of a term in a class stands for a simple idea: those terms that occur in many documents of a given class are good descriptors of that class.

DISCR

The **discriminative relevance** of a term in a class is based on the idea that a term is a good discriminator of a class if it tends to occur only in documents of that class.

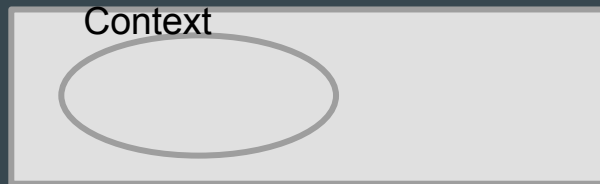
Proposed Technique

DESCR

The **descriptive relevance** of a term in a class stands for a simple idea: those terms that occur in many documents of a given class are good descriptors of that class.

DISCR

The **discriminative relevance** of a term in a class is based on the idea that a term is a good discriminator of a class if it tends to occur only in documents of that class



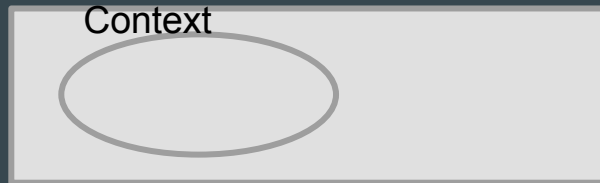
Proposed Technique

DESCR



The **descriptive relevance** of a term in a class stands for a simple idea: those terms that occur in many documents of a given class are good descriptors of that class.

DISCR



The **discriminative relevance** of a term in a class is based on the idea that a term is a good discriminator of a class if it tends to occur only in documents of that class

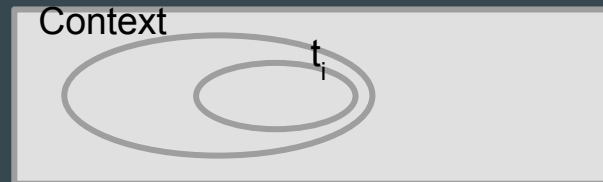
Proposed Technique

DESCR

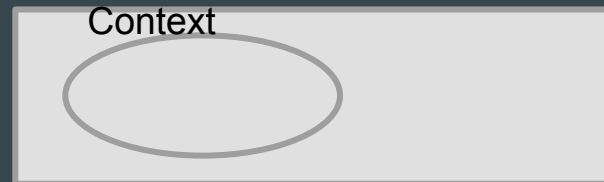


The **descriptive relevance** of a term in a class stands for a simple idea: those terms that occur in many documents of a given class are good descriptors of that class.

DISCR



The **discriminative relevance** of a term in a class is based on the idea that a term is a good discriminator of a class if it tends to occur only in documents of that class



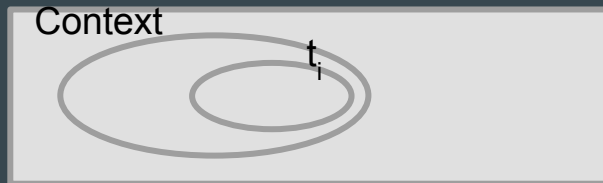
Proposed Technique

DESCR



The **descriptive relevance** of a term in a class stands for a simple idea: those terms that occur in many documents of a given class are good descriptors of that class.

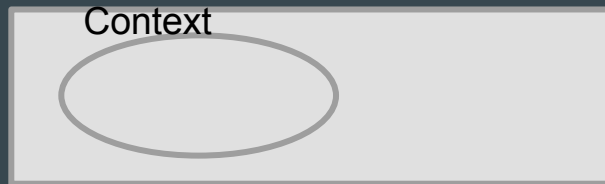
DISCR



The **discriminative relevance** of a term in a class is based on the idea that a term is a good discriminator of a class if it tends to occur only in documents of that class

FDD

$$\text{FDD}_{\beta}(t_i, c_k) = (1 + \beta^2) \frac{\text{DISCR}(t_i, c_k) \times \text{DESCR}(t_i, c_k)}{(\beta^2 \times \text{DISCR}(t_i, c_k)) + \text{DESCR}(t_i, c_k)}$$



Data Collection

Support The Guardian Subscribe Find a job Sign in / Register Search ▾

News **Opinion** **Sport** **Culture** **Lifestyle** More ▾

The Guardian International edition ▾

Data Collection

The screenshot displays the top navigation bar of The Guardian website. On the left, there is a 'Support The Guardian' button, followed by links for 'Subscribe', 'Find a job', 'Sign in / Register', and 'Search'. The main navigation menu includes 'News', 'Opinion', 'Sport', 'Culture', 'Lifestyle', and 'More'. The 'News' link is highlighted with a red underline. On the right side of the navigation bar, the 'The Guardian' logo is prominently displayed, with 'International edition' and a dropdown arrow to its right.

Below the navigation bar, a dark blue banner features a home icon and links for 'Get Started', 'Explore', 'Documentation', and 'Support'. The 'The Guardian Open Platform' logo is positioned in the top right corner of this banner. The main text of the banner reads 'Award-winning journalism Open to everyone' in large, bold, white font, with 'Open to everyone' being significantly larger. Below this, it states 'Access over 2 million pieces of content' in a smaller white font. A white circular icon with a downward-pointing chevron is located in the bottom left corner of the banner.

Data Collection

The screenshot shows the top navigation bar of The Guardian website. On the left, there is a 'Support The Guardian' button and links for 'Subscribe', 'Find a job', 'Sign in / Register', and 'Search'. The main navigation menu includes 'News', 'Opinion', 'Sport', 'Culture', 'Lifestyle', and 'More'. The 'News' link is highlighted with a red underline. On the right, the 'The Guardian' logo is displayed, along with a link for 'International edition'. Below the main navigation, a dark blue banner features a home icon, 'Get Started', 'Explore', 'Documentation', and 'Support' links. The banner text reads 'Award-winning journalism Open to everyone' and 'Access over 2 million pieces of content'. A white box with an orange border on the right side of the banner contains the text '20.840 News articles from 2013.' followed by a bulleted list: 'Politics.', 'Society.', 'Business.', and 'World news.'. A white downward arrow icon is located in the bottom left corner of the banner.

Support The Guardian

Subscribe Find a job Sign in / Register Search

News Opinion Sport Culture Lifestyle More

The Guardian International edition

Home Get Started Explore Documentation Support

The Guardian OpenPlatform

Award-winning journalism
Open to everyone

Access over 2 million pieces of content

20.840 News articles from 2013.

- Politics.
- Society.
- Business.
- World news.

Data Collection

The screenshot shows the top navigation bar of The Guardian website. On the left, there is a 'Support The Guardian' button, followed by links for 'Subscribe', 'Find a job', 'Sign in / Register', and 'Search'. The main navigation menu includes 'News', 'Opinion', 'Sport', 'Culture', 'Lifestyle', and 'More'. The Guardian logo is on the right, with 'International edition' next to it. Below the navigation bar, there is a dark blue banner with a home icon, 'Get Started', 'Explore', 'Documentation', and 'Support' links. The banner text reads 'Award-winning journalism Open to everyone' and 'Access over 2 million pieces of content'. A white circle with a downward arrow is in the bottom left. On the right side of the banner, there is a box with an orange border containing text about news articles from 2013, and a box with a red border containing text about manually labelled articles from January 2013.

Support The Guardian

Subscribe Find a job Sign in / Register Search

News Opinion Sport Culture Lifestyle More

The Guardian International edition

Home Get Started Explore Documentation Support

The Guardian OpenPlatform

Award-winning journalism
Open to everyone

Access over 2 million pieces of content

20.840 News articles from 2013.

- Politics.
- Society.
- Business.
- World news.

1.689 News articles from January 2013 were manually labelled by experts.

Validation I

- **Validation by User Study**

Terms were strategically selected from the dataset and manually scored by the users with a score between 0 and 5. We want to see the correlation between the human subject and our technique.

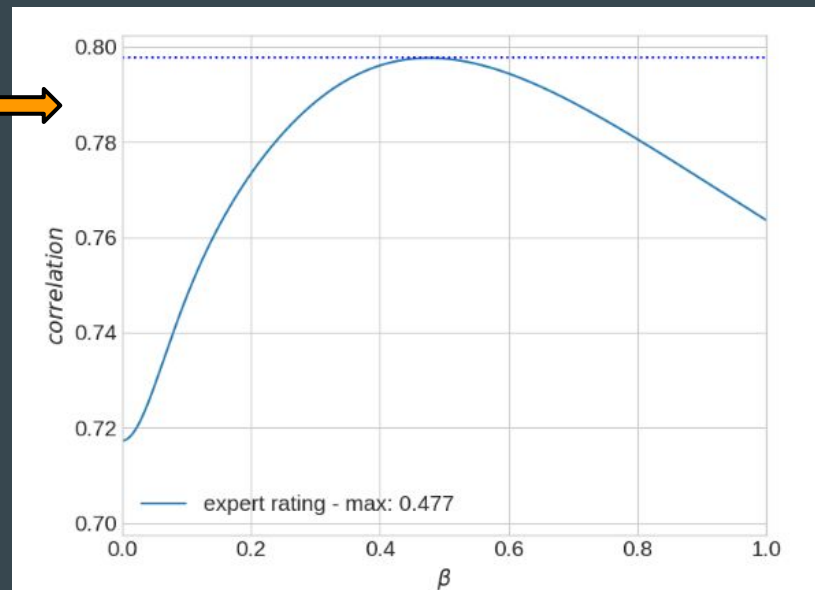
- A set of 50 terms for parameter tuning
- A set of 100 terms for validation

Validation I

- **Validation by User Study**

Terms were strategically selected from the dataset and manually scored by the users with a score between 0 and 5. We want to see the correlation between the human subject and our technique.

- A set of 50 terms for parameter tuning
- A set of 100 terms for validation



Validation I

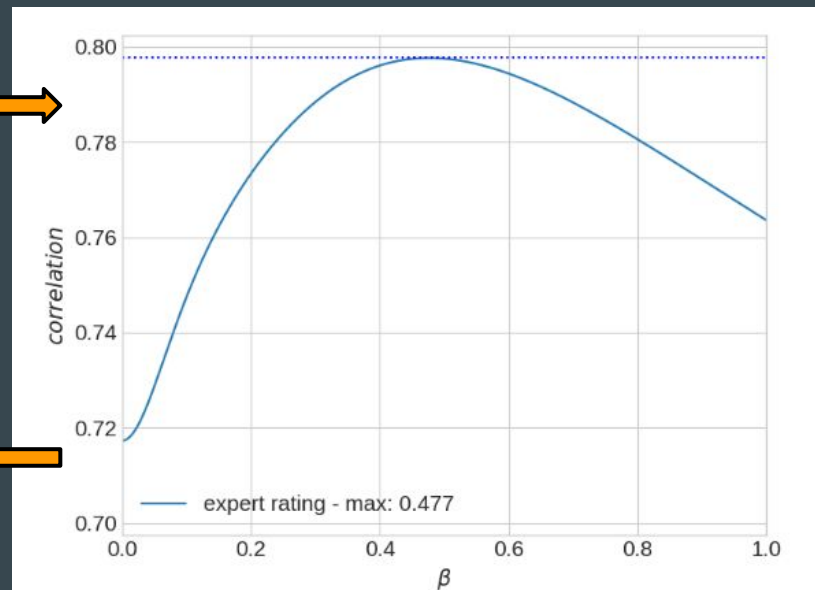
- **Validation by User Study**

Terms were strategically selected from the dataset and manually scored by the users with a score between 0 and 5. We want to see the correlation between the human subject and our technique.

- A set of 50 terms for parameter tuning
- A set of 100 terms for validation



non-experts and experts	non-experts and FDD _{0.477}	experts and FDD _{0.477}
$\mu = 0.80383, \sigma = 0.053205$	$\mu = 0.685598, \sigma = 0.054969$	$\mu = 0.752352, \sigma = 0.018904$

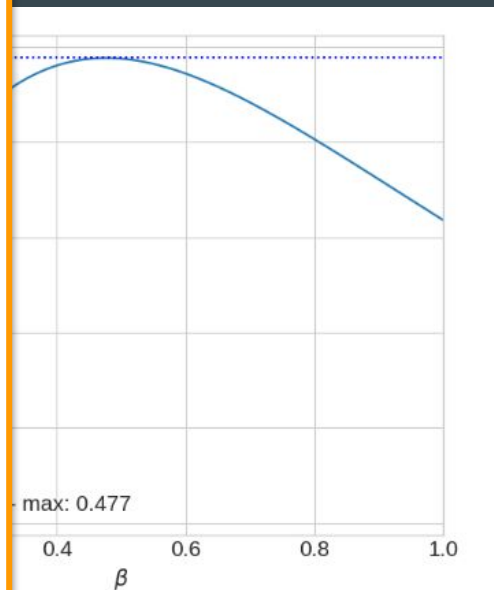


Validation I

- Validation by User Study

Terms were strategically selected from the dataset and manually scored by the
relation between the

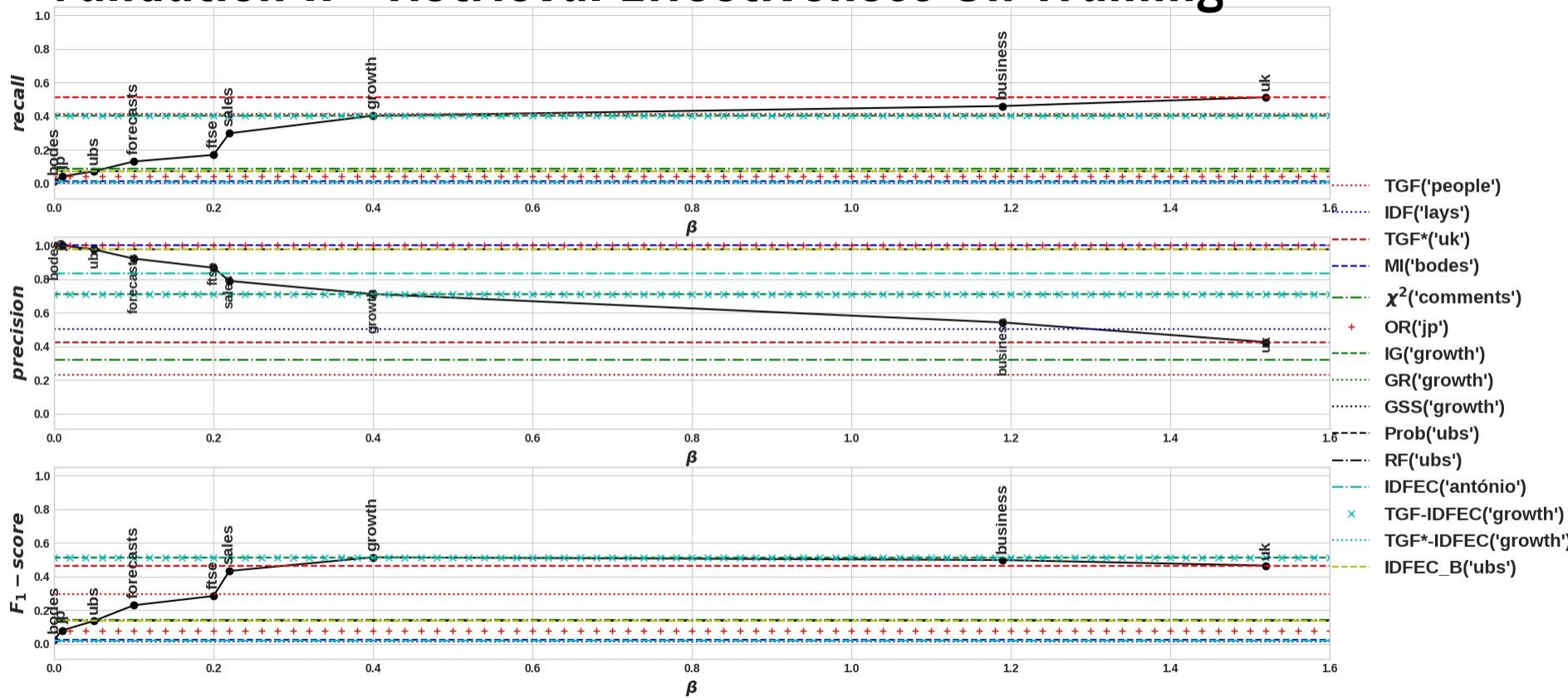
Method	non-expert (averaged)	expert (averaged)	non-expert and expert (averaged)
TGF	0.283553	0.365037	0.332324
IDF	-0.488816	-0.563704	-0.539138
TGF*	0.574110	0.642607	0.623198
MI	0.697053	0.659659	0.694604
χ^2	-0.164537	-0.087771	-0.128992
OR	0.432627	0.306599	0.378188
IG	0.663296	0.705736	0.701123
GR	0.663296	0.705736	0.701123
GSS	0.722761	0.757015	0.757807
Prob	0.654187	0.697007	0.691990
RF	0.472824	0.407394	0.450543
IDFEC	-0.226397	-0.325872	-0.283050
TGF-IDFEC	0.603975	0.676551	0.655882
TGF*-IDFEC	0.721871	0.774026	0.766110
IDFEC.B	-0.221061	-0.320304	-0.277466
DESCR	0.574110	0.642607	0.623198
DISCR	0.662481	0.610804	0.651848
FDD _{0.477}	0.735456	0.791969	0.782264



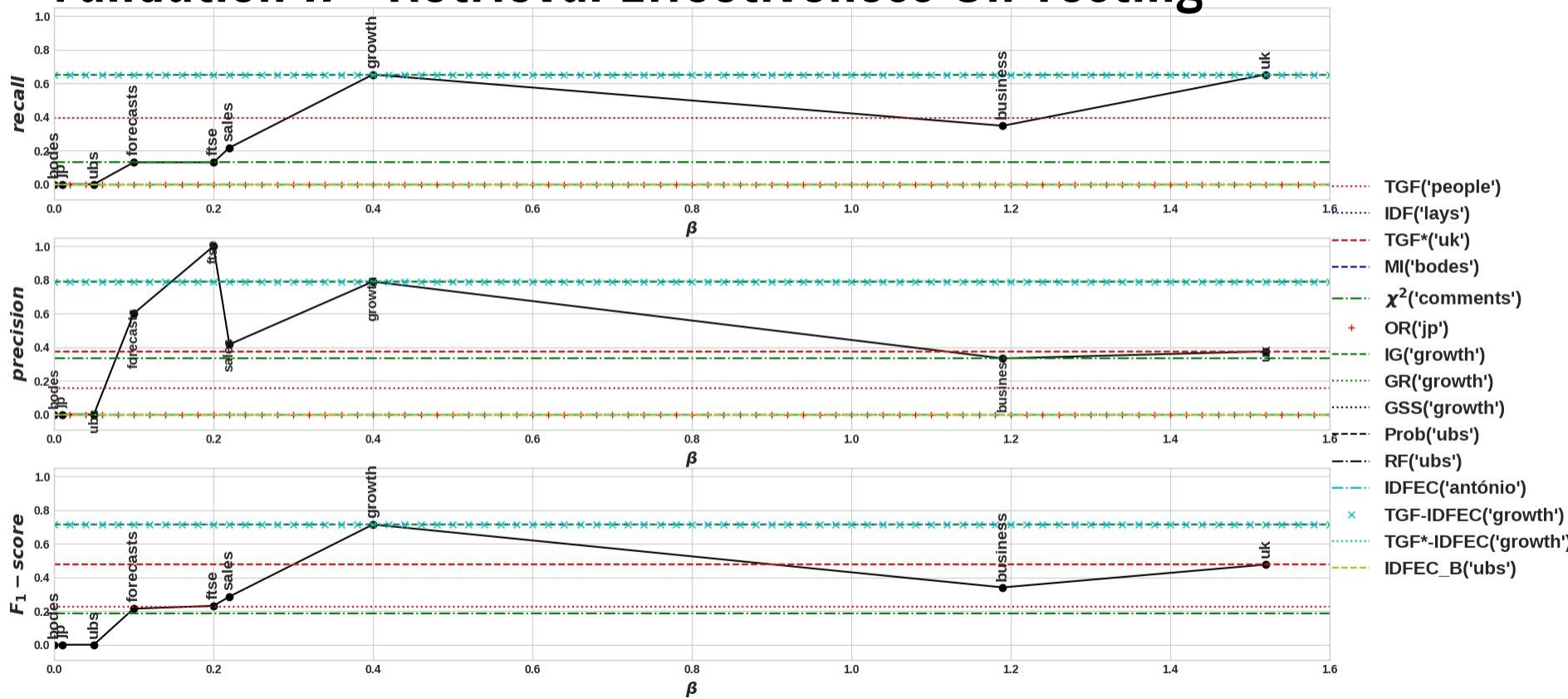
Validation II - Retrieval Effectiveness

- A reduced set consisting of 100 expert-labeled news articles (not included in the training set) was used as the validation set.
- The top-rated terms according to each technique were used as queries. The precision, recall, and f1-measure was reported.

Validation II - Retrieval Effectiveness On Training



Validation II - Retrieval Effectiveness On Testing



Conclusion and Future Work

- Good performance as an estimator of human subjects' relevance judgments.
- Good performance as a mechanism for selecting good query terms.

Conclusion and Future Work

- Good performance as an estimator of human subjects' relevance judgments.
- Good performance as a mechanism for selecting good query terms.
- Test FDD with more than two categories.
- Test different β for different datasets.
- A subsequent modeling step would be to identify different types of dependency relations between these variables (such as causal relations and close association).

Conclusion and Future Work

- Good p
- Good p
- Test FD
- Test dif
- A subse
relation



ments.

dependency
association).

THANK YOU
Questions?



mariano.maisonave@cs.uns.edu.ar