Pérez Téllez, F.; Cardiff, J.; Rosso, P.; Pinto Avendaño, DE.; Pinto Avendaño (2014). Weblog and short text feature extraction and impact on categorisation. Journal of Intelligent and Fuzzy Systems. 27(5):2529-2544. doi:10.3233/IFS-141227.

Weblog and short text feature extraction and impact on categorisation

Fernando Perez-Tellez[a,*], John Cardiff[a] , Paolo Rosso[b] and David Pinto[c]
[a]*Social Media Research Group, Institute of Technology Tallaght Dublin, Ireland*
[b]*NLE Lab.-PRHLT Research Center, Universitat Politècnica de València, Spain*
[c]*Faculty of Computer Science, Benemérita Universidad Autónoma de Puebla, Mexico*

Abstract. The characterisation and categorisation of weblogs and other short texts has become an important research theme in the areas of topic/trend detection, and pattern recognition, amongst others. The value of analysing and characterising short text is to understand and identify the features that can identify and distinguish them, thereby improving input to the classification process. In this research work, we analyse a large number of text features and establish which combinations are useful to discriminate between the different genres of short text. Having identified the most promising features, we then confirm our findings by performing the categorisation task using three approaches: the Gaussian and SVM classifiers and the K-means clustering algorithm. Several hundred combinations of features were analysed in order to identify the best combinations and the results confirmed the observations made. The novel aspect of our work is the detection of the best combination of individual metrics which are identified as potential features to be used for the categorisation process.

Keywords: Short Text Characterisation; Feature Extraction.

1. Introduction

The universe of weblog sites that share the opinions, commentaries or news on a particular topic, are often referred to as "*blogosphere*". The blogosphere brings to users the possibility of being a contributor of information instead of being only consumers of Internet information. We consider it important to characterise the blogosphere, and identify the principal features which distinguish weblog text among other types of short texts [17], as this allow us to perform more effectively clustering or topic detection tasks [22], [18]. For instance, if a text shows a high level of informality, a process of formalisation can be executed before the classification, task or if the text has short vocabulary size and low frequency terms an expansion or enriching process [8], [20] may help. In this sense, we expect that the assessment of the features of weblogs and short texts can impact the categorisation task in a positive manner.

Blogosphere texts have particular features by which they can be distinguished from other genres of short text which are identified and analysed in this paper. Firstly, we present an analysis using some metrics for the assessment of a set of corpus features that will provide us with insight into their nature based on a number of perspectives including shortness, lexical tags and formality level. We present the results of the experiments in which we analyse the features of a sample weblog corpora, and compare the results to other types of short texts extracted from more formal sources such as newsfeeds and scientific publications. In our experiments we have used four genres of short text: short news, computing scientific abstracts, weblogs and microblogs. The results show that we can isolate the most discriminating features for characterising short text genres, and we confirm that these can improve the categorisation task using two classifiers (SVM and Gaussian) and one clustering (K-Means) approach.

[*]Corresponding author. Fernando Perez-Tellez. Social Media Research Group, Institute of Technology Tallaght Dublin, Tallaght Dublin 24, Ireland.

We have focused our efforts on the identification of discriminating features and their application in improvement of the categorisation process. We have identified most relevant features for the categorisation task so we will be able to improve the task of organising this information in an automatic manner.

The rest of this work is organised as follows. Section 2 describes the related work. Section 3 describes the datasets used for the evaluation of features. Section 4 presents a discussion of features that may help in the identification of collections of short texts, such as microblogs, weblogs, short news and scientific abstracts. Section 5 describes our categorisation experiments and an analysis of the results. Section 6 presents the conclusions.

2. Related work

Characterisation of text has become increasingly more important to the research community in order to uncover features that typify the text type. There are some attempts which try to identify characteristics of weblogs that can be subsequently used to identify topics or categories automatically. In [12] the authors focus their efforts on detecting properties from core central weblogs. Their approach clusters weblogs according to their incoming links and their analysis is based on influence indicators, whereas in our case the focus is based on features contained in the text itself.

A research work based on traffic and communication patterns in the blogosphere is presented in [6]. This work characterises patterns based on the interaction between weblogs and their visitors, in contrast to our approach in which we are only interested in the identification of the relevant features based on the structure and linguistic features of the text.

The characterisation of short texts is discussed in [19] with the objective of detecting its impact on the clustering task. In this work, the authors focus their efforts on using classifier independent measures. They use short news collections in order to determine their quality (characteristics) in terms of four main categories of characterisation: domain broadness, class imbalance, stylometry and shortness. We are more interested in characterising different genres of text and doing a comparison among them rather than particular features within categories of a collection.

An attempt at characterisation of web content is also presented in [11], who present a research work that characterises reading level and topic distribution. This approach uses automatic text classifiers, based on reading difficulty and topic prediction, to estimate a type of probabilistic profile for important elements in the Web search such as users, web pages and queries. The objective of these profiles is to capture topic and reading level distributions which will be used in conjunction with log data in the characterisation and comparison of the users, web pages and queries. The characterisation used in this work is focused on the reading levels and topic distributions which will help to characterise the web content. We believe that these features are not suitable for categorising different types of short texts due to the fact that the documents may be offline with different topics and it will be difficult to estimate the distributions needed.

3. Datasets

In this section, we present the corpora used in our experiments. We have used what we called "reference" corpora which include collections of weblogs, microblogs, scientific abstracts and short news. We also have used "validation" corpora, as the name indicates; they are used for validation purposes and also contain all types of short text covered in this research work. We give a detailed description of these corpora in the following subsections.

3.1. Reference corpora

The datasets used in our experiments were chosen as representatives of their corresponding category (scientific abstracts, short news and weblogs) to determine relevant features of short text types (see Table 1). It is important to mention that we test our approach by comparing with different corpora of the same genres. We have selected short texts with different levels of formality: short news and scientific abstracts are by their nature formal texts while on the other hand, weblogs and microblogs are basically informal texts [16]. There are other characteristics that distinguish different types of short texts such as the length of the text (number of tokens); the percentage of out of vocabulary terms used in the text, the most used part-of-speech tags in text. In this research work, we use the words "term" and "token" with the same meaning and we define them as a string of non-blank characters usually, a word or atomic parse element.

The weblog corpora used in our experiments were selected to ensure that we had a representative sample of the blogosphere in general. Accordingly we constructed the following datasets for our principal experiments:

Table 1. Datasets used for feature evaluation.

	Dataset	Code	Type of Short Text	Number of Documents	No of Terms	Formal Text
Reference Corpora	Cicling 2002	(C02)	Computing-scientific abstracts	48	5112	Yes
	Reuters – Short news	(R8)	Short news	8158	2566548	Yes
	Weblogs – Narrow	(WbN)	Weblogs	25596	1511708	No
	Weblogs – Wide	(WbW)	Weblogs	48477	4695914	No
	Weblogs – Journal	(WbLJ)	Weblogs	29507	9072659	No
	Twitter data set (WePS)	(Mb)	Microblogs	34173	534656	No

Table 2. Datasets used for validation.

Datasets	Code	Type of Short Text	Number of Documents	No of Terms	Formal Text
Cicling 2005 Abstracts	C05	Scientific Abstract	29	2857	Yes
Easy Abstracts	EA	Scientific Abstract	48	5114	Yes
TSD 2010 Abstracts	TSD10	Scientific Abstract	20	2896	Yes
Text Mining Abstracts	TMA	Scientific Abstract	36	4873	Yes
Weblogs – Movies domain	WbM	Weblog Narrow	2876	160099	No
Weblog – Wide domain 2	WbW2	Weblog Wide	20510	4085811	No
Twitter subset 1 – General domain	TwS1	Microblog	950	4461	No
Twitter subset 2 – General domain	TwS2	Microblog	1150	2428	No
Twitter subset – Computing domain	TwC	Microblog	1200	9004	No
Micro4News	M4N	Short news	48	59944	Yes

- One wide-domain (WbW) and one narrow domain (WbN) dataset, both of which are constructed from collected from Yahoo Answers.
- A dataset of weblog posts by individuals (WbLJ), constructed from the well-known weblog site LiveJournal.
- A dataset based on Twitter (WePS), which is representative of the microblog genre.

The idea of having narrow and wide domain datasets is to establish if the proposed evaluation features are useful to discriminate between narrow and wide domains within weblog genre. Each of these are subsets of the ICWSM 2009 Spinn3r Blog Dataset [4]: a collection of 44 million weblog posts provided by Spinn3r.com. We selected these subsets because they are typical forums or personal journals which we are interested in analysing, and additionally we could find a large number of these documents in the dataset provided.

For the microblog category, the collection was obtained from the trial and training data sets of the Task 2 of the well recognised international competition known as WePS-3 evaluation campaign [1]. We would like to clarify that none of the collections were preprocessed in this case study because we wanted to keep as much information as possible from each of the collections.

The details of these datasets are shown in Table 1. In the scientific abstracts category we used the Cicling2002 (C02) collection which is made up of 48 documents from the computational linguistic domain. This corpus comprises the abstracts of papers appearing in the CICLing 2002 conference[1]. In terms of short news, we have used the well-known Reuters-21578 (R8) collection [13], which was collected from the Reuters newswire in 1987.

3.2. Validation corpora

In order to verify that the results of our experiments are not coincidental, we select a set of corresponding corpora which we term "validation corpora". We have selected a set of corpora of the same genres of short text to those already mentioned.

By repeating the experiments on these corpora, we can demonstrate that our findings are consistent.
Table 2 describes the ten datasets used for the validation of the results of the classification process. The microblog collections (TwS1, TwS2, TwC) were extracted semi-automatically using the Twitter4J API[2]. The collections TsS1 and TwS2 are wide domain and the TwC collection is considered narrow domain which includes computing related topics. The short news collection used is the M4N, which is available from the author's page [10] for research purposes. In the case of scientific abstracts, we have selected the EA corpus, C05 and TSD10[3] , which are computing related abstracts.

[1] http://www.cicling.org/
[2] http://twitter4j.org/
[3] http://www.tsdconference.org/tsd2010/

Table 3. Percentage of the collection with tokens of specific length.

Short Text Type	**Token Length (number of characters)**								
	One	Two	Three	Four	Five	Six	Seven	Eight	>8
Abstracts – C02	13.41% ± 0.26	15.89% ± 0.26	13.89% ± 0.32	10.03% ± 0.27	8.64% ± 0.28	5.28% ± 0.31	6.99% ± 0.26	7.09% ± 0.32	18.78%
Short news –R8	17.71% ± 0.04	14.89% ± 0.01	17.22% ± 0.02	13.56% ± 0.02	8.52% ± 0.02	7.33% ± 0.01	7.04% ± 0.01	5.31% ± 0.01	8.42%
Weblogs – WbN	19.98% ± 0.05	18.20% ± 0.03	18.69% ± 0.02	16.84% ± 0.02	8.25% ± 0.01	6.23% ± 0.01	5.01% ± 0.01	3.16% ± 0.01	3.64%
Weblogs – WbW	18.69% ± 0.01	18.91% ± 0.01	20.51% ± 0.01	17.67% ± 0.01	8.6% ± 0.01	5.85% ± 0.005	4.28% ± 0.006	2.37% ± 0.005	3.12%
Weblogs – WbLJ	22.40% ± 0.03	21.60% ± 0.04	19.60% ± 0.03	14.47% ± 0.02	6.47% ± 0.02	4.37% ± 0.01	4.08% ± 0.01	2.12% ± 0.01	4.89%
Microblogs – Mb	26.31% ± 0.04	13.40% ± 0.02	16.19% ± 0.02	13.83% ± 0.01	8.66% ± 0.02	7.73% ± 0.01	5.4% ± 0.01	3.22% ± 0.01	5.26%

These were manually constructed from the articles provided by the conferences websites, and TMA was manually constructed selecting random abstracts from articles related to the text mining area. Finally, the weblogs were automatically selected from Spinn3r corpus by filtering the weblogs with the category "movie" for WbM and selecting random categories for WbW2 subset.

4. Feature evaluation

We present a set of features and the hypotheses of how they may help us in finding discriminative information to be used in the characterisation process. This characterisation will lead us to detect important features that would be useful for relevant feature extraction from short text genres. We have divided the features into five sets based on intrinsic characteristics of the text, as described in the following subsections.

For each of the features presented in this section, the process of measuring the features was to select 70% of the documents of a collection randomly, then to compute the features and repeat the process twenty times to get the standard deviation and mean contained in each collection for each measure.

4.1. Length-based features

In this section, we present length-based features that we consider relevant to identify short texts of different types.

4.1.1. Token length

We define token length as $TL(C) = freq(t_n) / N$, where t_n is a token that contains n number of characters in it, N is the total number of tokens in a particular collection C.

We calculated the percentage of the collection which contains tokens with a particular number of characters. As can be seen in Table 3, the most frequent tokens are made up of one to four characters. We can see that microblogs contain the highest percentage of tokens with one character (26.31%). They also contain more punctuation symbols and signs than the rest of the short text types, which can have an impact in the categorisation task.

Scientific abstracts and short news contain longer tokens than the other genres of texts, tokens of seven and eight characters are more likely to appear in these types of texts. Short news (R8) contains a significant percentage of punctuation signs, articles and prepositions that cover most of the tokens with two and three characters.

Microblogs and weblogs showed consistent percentages i.e., no big variations over the lengths less than or equal to four, and as expected we obtained high percentages for short tokens due, among other factors, to slang or shortened words and informal text [24]. Based on this information we can say that if more than 50% of a corpus contains tokens of length less or equal to four, it is highly probable to be a corpus of microblogs. Weblogs reflect the same behaviour in both categories. The tokens with length less or equal to four can help to the identification of the category. In addition, if the percentage of tokens which are composed of more than eight characters is high, i.e., more than 15% the total of the corpus, they are likely to be abstracts; if the percentage of tokens is between the 5% and 10%, the text is probably short news; and finally, if the percentage of the corpus is lower they are found to be weblogs and microblogs.

4.1.2. Document length

We define Document Length as *DL(d)*, which is the number of tokens in document *d*. In Table 4, we present the average and the standard deviation of the length of documents in each collection. The number of tokens exhibits different characteristics in each kind of collection. Scientific abstracts typically are required to be less than 150 words and our results are in this range. The microblog corpus contains the smallest documents and the weblogs contain the next smallest

Table 4. Average document length (number of tokens).

Short Text Type	Document Length (tokens)
Abstracts – C02	106.38 ± 50.09
Short news –R8	134.97 ± 130.47
Weblogs – WbN	59.07 ± 44.35
Weblogs – WbW	96.77 ± 98.75
Weblogs - WbLJ	101.04 ± 90.56
Microblogs - Mb	15.63 ± 5.84

The short news contains the larger documents. Weblogs are also in the typical range of number of tokens per weblog post, less than 250 [3]. The standard deviation also shows values in the common ranges; for instance, the abstracts are usually limited up to 150 words but there are cases that they can be less than 100 words.

The short news category is less restricted in number of tokens, weblogs as expected are following the typical length and the standard deviation also shows to be in a valid range. In summary, the values obtained are valid according to the typical lengths of short texts such as in computing scientific abstracts which are limited to 150 words, microblogs are restricted to 140 characters, and weblogs and short news are usually around 200 words. Thus, we can say that document length is an indicator of the genre of text.

4.2. Vocabulary-based features

4.2.1. Lexical diversity in a document

This measure calculates the lexical richness of the text and lexical diversity (if the terms used throughout the text are diverse). The more varied a vocabulary of a text is, the higher the lexical diversity value will be. It is defined as follows:

*LD(d) = Length of document (d) / Vocabulary size (d) *100*

The principle of this measure is to pick the same number of tokens randomly from all text types, in our case we picked 4,000 tokens randomly from each corpora and then calculated the LD. Our expectation is that formal texts will have lower LD than weblogs and microblogs, due to the fact that formal texts are usually longer documents and allow authors to use a larger variety of tokens. Table 5 shows the degree of lexical information contained in each type of text. In the case of short news, the documents contain larger vocabulary than the scientific abstract documents but they have very similar value as narrow domain weblogs which is valid as short news also contain similar categories; on the other hand weblogs, in particular wide domain weblogs, have shown that this genre of text covers more diverse topics than the other weblogs. We can also see that the difference in vocabulary used is reflected in the results among weblogs and microblogs, the rate difference among vocabulary size and the number of tokens in microblogs is high. This is because bloggers in microblogs often use a wider variety of tokens, including those not included in a language lexicon, such as slangs and abbreviations.

Table 5. Lexical diversity.

Text Type	Lexical Diversity
Abstracts – C02	20.4 ± 0.70
Short news –R8	31.34 ± 1.54
Weblogs – WbN	30.47 ± 3.14
Weblogs – WbW	35.06 ± 1.79
Weblogs - WbLJ	31.04 ± 0.34
Microblogs - Mb	57.95 ± 1.79

4.2.2. Out-of-vocabulary tokens

Out-of-vocabulary (OOV) tokens are those absent from a particular lexicon. It is defined as:

Given a token w_x and a vocabulary $V=\{w_1,...,w_n\}$ if a token w_i $\neg\exists$ in V then w_i is a Out-Of-Vocabulary token.

In our experiments, we consider a token is OOV if it is not found in a general purpose English dictionary or in an English ontology. We consider it important to analyse this fact because there are new words continually coming into use that are not part of the current lexicons [23]. It is estimated in the same work that there are about 20,000 new words "born" every year, which means more than 50 per day. Our expectation is that more formal language such as abstracts contains tokens that can be found in a common English lexicon; however, in microblogs out-of-vocabulary terms: abbreviations, slang, misspellings and non-formal language are frequently used.

Table 7 shows the percentage of out-of-vocabulary tokens in our corpus. We have used a generic English dictionary included in GNU Linux systems which includes almost 100,000 words in English and the well-known WordNet thesaurus.

Table 6. Token occurrence (number of times used in the corpora) – percentage and standard deviation.

Short Text Type	**Token Occurrence (number of times used in the corpus)**					
	One	Two	Three	Four	Five	> Five
Abstracts – C02	56.83% ± 1.05	18.56% ± 0.97	7.93% ± 0.56	5.06% ± 0.45	2.60% ± 0.27	9.02% ± 1.56
Short news –R8	35.37% ± 0.18	13.80% ± 0.17	7.67% ± 0.14	5.17% ± 0.09	3.61% ± 0.08	34.38% ± 1.03
Weblogs – WbN	55.51% ± 0.28	13.67% ± 0.19	6.10% ± 0.07	3.78% ± 0.09	2.4% ± 0.09	18.54% ± 1.69
Weblogs – WbW	53.45% ± 0.20	13.89% ± 0.18	6.36% ± 0.10	3.89% ± 0.06	2.4% ± 0.05	20.01% ± 1.78
Weblogs - WbLJ	50.83% ± 0.16	14.81% ± 0.13	6.21% ± 0.05	4.98% ± 0.06	2.65% ± 0.03	20.52% ± 2.03
Microblogs - Mb	65.07% ± 0.12	12.13% ± 0.14	5.10% ± 0.08	3.02% ± 0.06	2.04% ± 0.05	12.64% ± 2.15

The results show a large difference between the microblogs (31.19% for the English dictionary and 40.01% for the lexical database) and the rest of texts. In addition, the restriction of size for microblogs forces users to use abbreviations in their text. As expected more formal texts contain less out-of-vocabulary tokens. Short texts which tend to have more than 30% of the total corpus OOV are very probably microblog texts.

Table 7. Out-of-vocabulary tokens.

Text Type	**Out-of-Vocabulary English Lexicon**	**Out-of Vocabulary Wordnet**
Abstracts – C02	4.27% ± 0.38	5.11%
Short news –R8	8.24% ± 0.03	10.85%
Weblogs – WbN	10.13% ± 0.07	13.56%
Weblogs – WbW	7.4% ± 0.02	12.25%
Weblogs - WbLJ	17.72% ± 0.10	20.02%
Microblogs - Mb	31.19 % ± 0.06	40.01%

For weblogs, the fact that the narrow collection WbN contains more technical terms than wide domain is reflected in the results by having more OOV terms, i.e., generic resources do not contain high level of technical vocabulary. However, the resources used do not contain specialised terms, and as a consequence these kinds of terms are considered as OOV. Based on this, we consider this feature valuable for differentiating general purpose weblogs from the technical weblogs.

4.2.3. Frequency of token occurrence

This aspect measures the percentage of the total corpus with a given token occurrence frequency. We define this measure as:

$$FT(i,j) = O_{Wi} / N_j$$

where O_{Wi} is the number of occurrences of the token Wi, $0 < O_{Wi} < 6$ and N_j is the number of tokens in corpus j. We hypothesise that text with a high percentage (more than 60%) of token occurrence equal to one are likely to be microblog text, hence the tokens used in microblogs are not commonly repeated in the same text.

We also expect that the percentage of token occurrence greater than five will be higher for scientific abstracts than the rest of the texts because they contain specialised terms. This information will help in determining the richness of the text.

Table 6 presents the token occurrence for each corpus the values representing the percentage of the total collection. Our results show that tokens that appear only once in the collection cover a large portion of the corpora with more than 55% in scientific abstracts, microblogs and weblogs. We find that microblogs have the highest percentage of tokens that appear only once in the collection with 65% of the tokens. This is generated because punctuation signs used as ":)" and ":P" among others, misspelled words and abbreviations occur frequently in microblogs. Scientific abstracts have a relatively high percentage of tokens that appear once (56.8%) in the collection and appear twice (18.56%). This is because of the specific vocabulary used in this collection. Short news does not display a high level of tokens that appear once (35.5%) because in most of the news, authors tend to re-use the same tokens in their story due to the fact that the categories are very similar in the R8 collection. The WbN and WbW domains of weblogs show the same pattern in this respect. The microblogs corpus contains 65% of the whole collection with tokens that appear only once, and then it is because the variety of vocabulary and out-of-vocabulary terms used in the microblogs is larger than in formal texts. Short text with more than 60% of the token occurrence equal to one is very likely to be microblog and text with more than 30% of the collection with token occurrence greater than five is probable to be scientific abstracts.

In addition, Table 6 also presents the token occurrence greater than five, and as it was expected there is a high percentage of R8 (more than 30%) which contain tokens that occur more than five times, it is because the formality of the news and very narrow domain tend to use the same terminology. We consider that the frequency of token occurrence equal to one can help the identification of the collections, in particular for Mb and C02 collections and on the opposite side the R8 collection has low value in the token occurrence equal to one.

4.3. Part-of-speech-based features

The part-of-speech category (POS) categories we evaluated are shown in Table 8.

Table 8. Part-of-speech tags.

Tag	Part-Of-Speech
NN	Common noun, singular or mass
NNS	Common noun, plural
JJ	Adjective, comparative
IN	Conjunction, coordinating
DT	Determiner, it includes articles
PRP	Personal pronoun
:	Other characters or strings
.	Punctuation signs
CD	Cardinal number
VBN	Participle, past

Following the outcome of the analysis conducted by [9] we expect that more nouns, adjectives and non-deictic words are commonly used in formal writings and on the other hand, personal pronouns, the use of deictic words punctuation marks and shortening words are used in informal texts such as microblogs and weblogs and on this basis we may be able to distinguish between text corpora.

In Table 9 we can see that scientific abstracts exhibit high levels of nouns and adjectives, followed by conjunction and determiners because they are mostly used in descriptions and explanations; on the contrary, abstracts have shown low levels of CD (cardinal numbers) in comparison with the rest of the corpora with only 1.16%. In short news, nouns are predominant (23.85%) but not at the same level as in microblogs (32.16%), in which nouns seem to be the most frequent content POS tag used in this collection. It is well-known that nouns are meaningful words in the text as they are part of the category known as "content words" or "lexical words" [21]. Weblogs contain relatively high percentages of pronouns (PRP) with 8% for WbN, 11% of the WbW corpus and 13% of the WbLJ corpus. This is in agreement with [9] who state that high pronouns usage is an indicator of an informal writing style. Based on the results obtained the POS tags that can be considered for the classification step are nouns (NN, NNS), determiners (DT), pronouns (PRP), and the category "other characters or strings" (":").

4.4. Formality-based features

The style of writing results from the fact that different people express themselves in different ways and that the same person would express the same idea differently when addressing different audiences [2]. Therefore, we believe it should be possible to automatically recognise the distinction between formal and informal manner of writing expression and use this information in the identification of genre of text. We have used two different formality measures from the literature, the more traditional measure that was introduced by [9] and the relatively new measure presented in [16]. We also present a new measure, developed by ourselves, which is based on the [9] approach.

4.4.1. Heylighen & Dewaele measure

We have decided to use a measure in order to obtain informality levels in the different kind of texts in discussion. This measure is based on POS tags and uses the POS annotations and their frequency of appearance. In [9] the authors describe formality/contextuality as "dimension of variation between linguistic expressions". In the same work they propose a metric Eq. (1) to measure the level of formality of text.

$$F0 = [(\text{noun freq.} + \text{adjective freq.} + \text{preposition freq.} + \text{article freq.} - \text{pronoun freq.} - \text{verb freq.} - \text{adverb freq.} - \text{interjection freq.}) + 100\ ]/\ 2 \quad (1)$$

As this measure was not designed to evaluate these new genres of texts such as weblogs and microblogs, it does not consider emoticons, abbreviations or the shortening of text that is commonly found in informal text. The basis of *F0* formula is lexical analysis which at the time of the proposal no genres of text such weblogs and microblogs were defined and widely used.

4.4.2. Modified Heylighen & Dewaele measure

We have modified the original measure in order to evaluate the formality of the different types of short texts covered in this research. The original formula was modified with the purpose of giving more importance to the "content" words (formality) and seeing the impact of the "functional" words of the text that in theory according to [9] are the less formal. On this basis, the lower the value the more informal the text is. The logic behind this new formula Eq. (2) is to give more weight to the content words and incorporate new categories such as strings that represent emoticons or a shortening of a word for example "10q" instead of using "Thank you" or the simple ":)" as "I am happy" or "I like" which cannot be identified in any POS category. We also considered it important to take into account the presence of other symbols within the text such as emoticons, abbreviations and punctuation marks.

Table 9. Most used part-of-speech categories.

Short Text Type	**Part-Of-Speech tags**									
	NN	NNS	JJ	IN	DT	PRP	:	.	CD	VBN
Abstracts – C02	21.52% ± 0.34	9.70% ± 0.25	9.47% ± 0.33	12.69% ± 0.29	11.38% ± 0.26	1.18% ± 0.08	3.37% ± 0.21	3.98% ± 0.09	1.16% ± 0.14	3.90% ± 0.21
Short news – R8	23.85% ± 0.02	8.84% ± 0.01	5.70% ± 0.013	9.81% ± 0.01	7.41% ± 0.01	1.49% ± 0.008	1.88% ± 0.01	5.34% ± 0.01	8.20% ± 0.05	2.46% ± 0.07
Weblogs – WbN	19.31% ± 0.04	3.97% ± 0.01	3.92% ± 0.008	7.28% ± 0.01	7.14% ± 0.01	8% ± 0.02	5.42% ± 0.04	4.97% ± 0.01	3.65% ± 0.02	1.33% ± 0.006
Weblogs – WbW	14.90% ± 0.02	3.85% ± 0.006	4.14% ± 0.007	8.30% ± 0.01	5.75% ± 0.007	11.09% ± 0.01	5.15% ± 0.02	4.55% ± 0.01	3.15% ± 0.01	1.39% ± 0.004
Weblogs - WbLJ	20.62% ± 0.03	4.01% ± 0.01	4.20% ± 0.001	8.42% ± 0.02	7.32 % ± 0.003	13.09% ± 0.001	7.01% ± 0.03	5.11% ± 0.01	2.98% ± 0.01	1.35% ± 0.005
Microblogs – Mb	32.16% ± 0.03	5.03% ± 0.01	3.70% ± 0.01	5.45% ± 0.01	3.54% ± 0.01	2.65% ± 0.01	13.76% ± 0.03	6.08% ± 0.01	4.29% ± 0.02	0.79% ± 0.006

We define a new category "other strings" which includes these new characteristics of text (emoticons, abbreviations and punctuation marks). These characteristics are usually employed in weblogs and microblogs and they are indicators of informal text. In addition, this modification helped to identify easily the boundaries between different short text types.

$$F1 = [(\text{noun freq.} + \text{adjective freq.} + \text{preposition freq.} + \text{article freq.} - \text{pronoun freq.} - \text{verb freq.} - \text{adverb freq.} - \text{interjection freq.} - \text{other strings freq}) / (\text{noun freq.} + \text{adjective freq.} + \text{preposition freq.} + \text{article freq.}) + 1] \quad (2)$$

According to this proposal, nouns, adjectives, articles and prepositions are more likely to be found in formal styles, whereas pronouns, adverbs, verbs and interjections are more probable in contextual styles.

The outcomes of our evaluation for *F0* and *F1* are shown in Table 11. Apart from nouns, we have observed a high percentage of interjections, pronouns and adverbs which are more likely to appear in subsets of microblogs and weblogs. With the *F1* measure the higher the value, the higher the formality of the text.

In Table 11, we present the evaluation of *F0* and *F1* formality measures, *F0* is the original measure proposed by the authors, it gives very similar values of the formality measure of the collections but on the other hand, *F1* is better adjusted to evaluate short texts thus is easier to identify the level of formality. We also see that *F1* can give us a better discrimination between the collections (see Figure 1).

We would like to mention that in the case of WbN and WbW collections, we did not find the same percentage of nouns and punctuation/other signs as in microblogs. The articles, verbs and adverbs are significant for WbN and WbN which decrease the value of informality of *F1*. We can also see that verbs, adverbs part-of-speech categories and "other strings" (emoticons, abbreviations and shortening of words) are highly exploited by many microblog users. On the other hand, weblogs contain more personal pronouns compared with microblogs reflected in their 0.34 and 0.19 of formality value. Abstracts and short news showed that nouns are more frequently used; they also have a low level of pronouns, interjections and adverbs. Based on this information we can say that formal short text are in the range of greater than 0.7 in *F1* and informal short text can be considered in the range of less than 0.35 for the same metric. From these results, we can see that the *F1* measure is a more successful as indicator of formality.

4.4.3. I-measure

I-measure is another measure for the formality of text that is based on POS tags. This metric was proposed in [16] to obtain an indicator of the informal characteristics used in a text such as misspelled words, level of interjections and emoticons, and be able to evaluate the level formality of a text. This metric has been applied to evaluate a subset of social media text included in the collection of the "Fundación Barcelona Media"[4]. This *i-measure* is defined as:

$$I\text{-}Measure = (\text{Wrong-typed Words freq.} + \text{Interjections freq.} + \text{Emoticon freq.}) * 100$$

I-measure appeared due to the fact that in Web 2.0 exists very different text types such as microblogs, chats, comments in social media and each with different characteristics. From this point the necessity of measuring the quantity of informal factors such as emoticons, colloquial expressions, slang or chat-style abbreviations became important for the analysis of text and for its categorisation.

[4] Fundacion Barcelona Media, 2009, Caw 2.0 training datasets. http://caw22.barcelonamedia.org/node/7

In Table 11, we also show the level of informality of each collection (*i-measure*); it is important to mention that the "wrong-typed words", "interjections" and "emoticon" frequencies were calculated in this case for each document in the respective collection and the average was used. The results show, as expected that the value is higher for microblogs (with 31.33 *i-measure* value). The values from abstracts and short news are low as expected because no misspelling or emoticons should be present but in particular short news showed close values to weblogs wide domain, which is basically because they are in the wide domain, i.e., the topics discussed in both kinds of text are diverse and we find the use of interjections, in other words, they share the same rate of frequency of interjections and not values in wrong-typed words and emoticons frequencies.

We have found this metric valuable because we can clearly find well-defined boundaries in all genre of text. Even though we cannot give exact ranges for this metric due to the fact that the *i-measure* metric is not normalised we can clearly recognise if a short text belongs to a particular category based on this measure values which have clear boundaries among genre of short text and a categorisation algorithm would be able to group text of the same type.

Finally, we would like to compare the two formality measures *F1* and *i-measure* because we have found that *F0* does not incorporate any "new" informal characteristics of text used in the blogosphere such as slang, chat-style abbreviations, emoticons or colloquial expressions. We consider that *F1* produces better indicators of informality because this measure includes POS categories which have been identified as more popular for informal text. *I-measure* gives us good boundaries to group the same type of text together but it is more focused on text produced in the social media platforms as which was the purpose of the authors. Even though the *i-measure* does not provide the best well-defined boundaries for the formality of a text, we still consider it important to take into account to see the benefit that can provide to the categorisation process.

4.5. Statistical-based feature: Language models

We anticipate that the language model (a probability distribution over entire sentences or texts) may be different for each kind of short text, for instance we believe that scientific abstracts may present better language models than the other types because the domain is restricted and the distribution can model more precisely the language used in this genre of text. We have used perplexity [15] in order to compare them.

The aim of this measure is to compute the predictability of text of the different genre of short text. The hypothesis states that perplexity can be used as indicator to determine the genre of text; formal text will have lower value than informal text.

Table 10. Perplexity of the Language Models.

Short Text Type	**Perplexity**
Abstracts – C02	162.86 ± 10.20
Short news –R8	218.15 ± 2.34
Weblogs – WbN	232.57 ± 3.31
Weblogs – WbW	197.79 ± 2.06
Weblogs - WbLJ	229.81 ± 4.30
Microblogs - Mb	478.79 ± 10.58

Table 10 presents the perplexity of each collection. We can see that it is difficult to determine the predictability of microblog text due to its diversity. We also found it difficult to use this feature for distinguishing between short news and weblogs collections due to the fact that these genres of texts contain different topics and they are less predictable. The more formal the text is, the more predictable it may be; which is because the weblogs in all the cases and microblogs writers do not use a specific style and the way of writing may change from one author to another. This measure also depends on the broadness of the collection, i.e., how broad the topics contained in the collection are. The more topics covered in a collection is more difficult to determine the predictability of the text. The valuable information gathered from this feature is the confirmation that the difference between abstracts and microblogs is clear, the perplexity value from abstracts is less half of microblog value.

4.6. Summary of findings

We have evaluated the features of different genres of short texts and we have identified that some of the measures will benefit the classification process based on the well-formed boundaries between texts features. In the *Length-based Features*, we found an important value in analysing the length of tokens and documents from each collection. Authors use different length of tokens and documents to express their ideas, depending on the category of the text.

This fact can be originated by the limitation of the platform, for instance in microblogs texts will contain a distinct number of tokens and the tokens used are different in length compared to longer texts such as scientific abstracts or short news.

Table 11. Formality measures for each genre of text.

Short Text Type	Heylighen & Dewaele (F0)	Modified Heylighen & Dewaele (F1)	I-measure Value
Abstracts – C02	50.28 ± 0.01	0.81 ± 0.05	4.18 ± 0.27
Short news –R8	50.21 ± 0.03	0.74 ± 0.06	6.23 ± 0.03
Weblogs – WbN	50.09 ± 0.05	0.341 ± 0.01	10.20 ± 0.05
Weblogs – WbW	50.04 ± 0.07	0.105 ± 0.01	7.446 ± 0.02
Weblogs – WbLJ	50.06 ± 0.06	0.216 ± 0.02	8.56 ± 0.01
Microblogs - Mb	50.19 ± 0.09	0.199 ± 0.01	31.332 ± 0.07

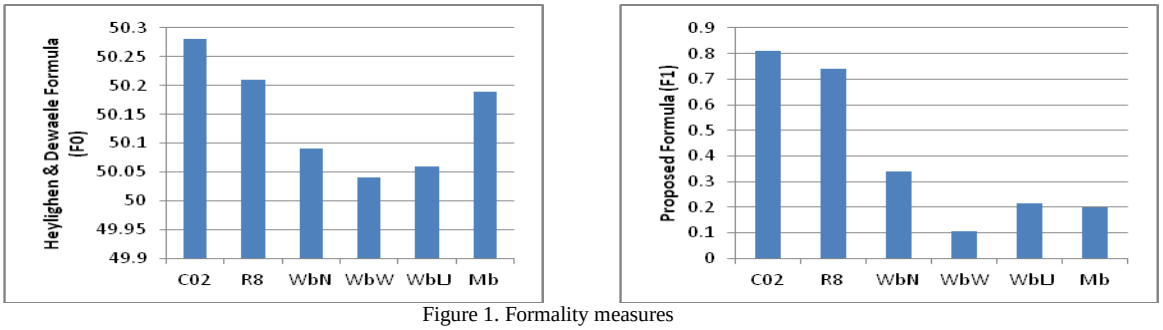

Figure 1. Formality measures

Vocabulary-based Features have measured the richness of the text based on the vocabulary size and the document length of a collection, the percentage of the corpus in which tokens are used only a few times over the whole collection and the percentage of terms that are not found in the English lexicon. The value of this set of features is to explore the vocabulary characteristics in each collection. We found different patterns for each feature measure in each type of short text. For instance, if a collection contains large tokens (greater than or equal to eight characters), it is likely to be abstract text, if the percentage is between 5% and 10% the text may be from short news and lower than 5% they would be weblogs or microblogs.

Regarding the *Part-of-speech-based Feature*, we have analysed the POS tags most commonly used in each type of text. Authors use different parts of speech depending on the genre of text that they are writing. In this category, we found that nouns (NN, NNS), determiners (DT), pronouns and the category "other characters or strings" are good candidates to be used as indicators for short text types.

Under *Formality-based Features*, we measured the level of formality of the corpora for every genre of short text. We tested a traditional formula and we adapted it to be able to apply it to the new genres of text generated in the blogosphere. We also tested another formality metric which showed to be able to differentiate among the genres of text. We found that the best discriminative information was provided by the measure we propose, *F1*.

Finally, the *Statistical-based Feature* we used, perplexity, helps us to determine the predictability of the collections. We found that the predictability of a language model is directly related to the broadness and formality of the text, i.e., the more formal and narrower the domain is, more predictable it may be.

We are interested in verifying if the discriminatory features presented in this work can be used to effectively categorise a given text. We also found it difficult to differentiate between narrow and wide domain weblogs, as the values obtained for many features are very similar, however as we mentioned some features can be employed by a categorisation process to identify the genre of text**.** We have used different combination of these features in order to detect which ones provide better discriminative information to the classification process. In Section 5, we provide a detailed evaluation of an unsupervised and two supervised approaches to categorisation of the texts, based on the features we have analysed.

5. Impact of the features in categorising short text genres

In this section, we focus our efforts on analysing the features presented and evaluated in Section 4. Each feature will be represented in a vector which will be the input of a clustering algorithm and two classification algorithms with the objective of detecting which features produce acceptable results. The clustering algorithm used is the well-known K-means algorithm and the classification approaches used are Gaussian and the SVM classifiers. Finally, we compare our results with other works.

5.1. Description of experiments

In this section, we show the performance of the categorisation approaches using different combinations of the features discussed above. Initially, we use the corpora described in Table 1 and subsequently repeat the experiments using the validation datasets (Table 2).

The extraction of the features was performed over all collections and then the features were used in the categorisation process. The idea is to analyse the contribution of the features in the improvement of results when these algorithms are applied to them, and we expect to identify the best features for a clustering algorithm and show improvement in the performance.

We define a vector as the representation of a collection formed with the features discussed previously. The feature-vectors are a compound of thirty one dimensions, in which the first thirty dimensions correspond to the features and the last dimension describes the category of the collection (abstract, short news, microblog or weblog).

The purpose of the experiments is to identify the set of features for each of the categorisation algorithms which yield the best results. Accordingly we performed the categorisation for every combination of each of the nine features which resulted in 512 different executions for each feature combination.

The input for the Gaussian and the SVM classifiers is the feature-vectors generated by the reference corpora which are used as the training set in the classification processes. The validation corpora are used as test sets in the classification process. Each vector contains the values of the features already discussed, and the output of this process is to label the evaluation corpora with the short text genre assigned by the classification task.

The process of building the training set for the Gaussian and SVM classifiers with feature-vectors from the reference corpora was repeated twenty times with different documents for each collection in the reference corpora; in each execution a vector was generated from random documents in the collection. These were used in the categorisation process as the training set. For the testing set a vector of each validation corpus was generated. The output is the category labels (scientific abstracts, short news, weblogs or microblogs) for each testing collection.

The vectors are constructed from the reference and validation corpora for the clustering approach. We have used the K-means algorithm. The process of building the vectors also starts with the extraction of features, but in this case using both set of collections, the reference and validation datasets, one vector is generated per collection. After the feature extraction, the vectors generated from this step are sent to the clustering algorithm and four different groups are generated.

In order to validate our results and to obtain more generic conclusions, we have shifted ten times the reference corpora with the collections in the validation corpora, so as to create ten different combinations among the collections, i.e. we replaced the reference collections with collections in the validation datasets and executed the experiment as before.

5.2. Results and analysis

In this section, we present the results obtained from the experiments. We present the average of these ten executions with different training and testing datasets for the Gaussian and SVM classifiers and different representative collections of each category are used for K-means. Furthermore, we have carried out the additional experiments with the reference and validation collections in order to verify that our results were not a coincidence.

5.2.1. Gaussian classifier

In Table 12, we show a selection of the best results obtained using the Gaussian classifier. It can be seen that the best result achieved was 0.67 (F-measure), which was obtained using the feature combinations shown. These values represent the average from ten executions with different datasets. The lexical diversity and document length features are shown to be important in this approach as they are consistently appearing as good features. However the formality features (F1 and iM) and out-of-vocabulary feature also demonstrate significant impact. Perplexity gave good improvement when it is used together with the features just mentioned. The level of formality is also an important factor to be taken into account for distinguishing among these types of texts indicating that authors tend to use different styles of writing (formality level) depending on the type of text they are writing.

This classifier achieves good performance by using LD, DL features and the formality measures F1 and iM. In addition, the OOV and TL features consistently provide discriminatory information to the classification task. In conclusion, document and token length, lexical diversity and formality measures are important factors for the classification process based

on the SVM classifier as writers use different terminology and levels of formality according to the type of text they are producing.

Table 12. Best combinations of features for Gaussian classifier.

ID	Relevant Features for Gausssian Classifier	F-measure
G1	LD, TL, DL, F1, PPL, iM	0.67
G2	LD, OOV, TL; DL, F1, PPL	0.67
G3	LD, FT, OOV, DL, F1	0.67
G4	LD, OOV, TL, DL, F1, iM	0.66
G5	LD, OOV, DL, F1, iM	0.66
G6	LD, OOV, TL, iM	0.65
G7	LD, F1, iM	0.63
G8	LD, TL, DL, PPL, iM	0.63
G9	LD, OOV, F1	0.59
G10	LD, FT, DL, F1, iM	0.59
G11	LD, DL, F1, POS, iM	0.57
G12	LD, FT, OOV, DL, POS	0.57

5.2.2. SVM classifier

Table 13 shows us the best results obtained using the SVM classifier by using specific combinations of features. The best F-measure value obtained using the SVM classifier is 0.66 which is very similar from the one obtained with the Gaussian classifier.

Table 13. Best combinations of features for SVM classifier.

ID	Relevant Features for SVM Classifier	F-measure
S1	LD, DL, F1, iM	0.66
S2	LD, OOV, DL, F1, PPL, iM	0.66
S3	LD, OOV, TL, DL, F1, iM	0.66
S4	LD, OOV, TL, DL, F1	0.66
S5	LD, OOV, DL, F1, iM	0.65
S6	OOV, DL, F1, iM	0.65
S7	LD, F1, iM	0.65
S8	LD, OOV, TL, DL, iM	0.62
S9	LD, OOV, F1	0.62
S10	LD, DL,F1, iM	0.61
S11	LD,OOV,TL,DL	0.61
S12	LD, OOV, DL, F1, POS	0.60

5.2.3. K-means clustering

Table 14 presents the best results (average) of executing the 512 combinations of features with the K-means algorithm. For the results presented, the parameter k was set as 4 but we have performed the experiments with k from 2 to 6 and the best results were achieved with $k = 4$, corresponding to the four groups of short text (scientific abstracts, short news, weblogs and microblogs). The results we show in Table 14, where we can see the best value obtained is 0.63 F-measure which is achieved when LD, FT, TL, DL and iM either F1 features were used. We have seen that LD and DL are consistent in all cases for achieving good results in the clustering process. Moreover, the formality measures F1 and iM are also important in most cases.

Table 14. Best combination of features for K-means algorithm.

ID	Relevant Features for K-means	F-measure
K1	LD, FT, TL, DL, iM	0.63
K2	LD, FT, TL, DL, F1	0.63
K3	LD, OOV, TL, DL, F1	0.61
K4	LD, FT, TL, DL	0.61
K5	LD, DL, F1, iM	0.60
K6	LD, DL, F1, POS, iM	0.60
K7	LD, DL,POS, iM	0.58
K8	LD, DL, F1, POS	0.58
K9	LD, OOV, TL, DL, POS, iM	0.58
K10	LD, OOV, TL, DL, POS	0.55
K11	LD, DL,POS	0.54
K12	LD, OOV, DL, POS, iM	0.54

5.3. Critique and analysis of approaches

We have seen that the lexical diversity and document length are important factors to be considered for differentiating among short text types. The formality measures can help in creating well-defined boundaries among the categories due to the fact that short text writers use different language according to the platform used and the audience directed to. FT did not emerge as a relevant feature, as it did not appear as a feature in the best SVM classification results and made almost no impact in the Gaussian classification. We have seen small impact in the results when POS feature is used in all the approaches.

The clustering algorithm has shown good performance, which is comparable with the classifiers. Even though it did not get the best result, this was expected because it is an unsupervised approach, and it proved to be an adequate option when a training set is not available. The unsupervised approach minimises relatively easily the objective function, using the features presented above appropriately. The need for training datasets can be a drawback in some cases for particular domains and for such cases, the clustering approach may be the best option.

We have applied several combinations of features to the three categorisation approaches and identified the best features per each categorisation task from Table 12, 13 and 14.

- The Gaussian classifier has shown to work better with the following features: LD, TL, DL, F1, PPL and iM, and also when iM is substituted for OOV. Another combination which has shown good results is: LD, FT, OOV, DL and F1.
- The SVM classifier performed well with the following features: LD, DL, F1 and iM. The classi-

fier also showed good result when OOV and PPL features were incorporated.
- The K-means clustering algorithm has produced the best results using the following features: LD, FT, TL, DL, iM and when substituting iM for F1.

5.4. Comparison with other work

We have focused our efforts in order to identify discriminative features in different categories of short text (scientific abstracts, short news, weblogs and microblogs) for categorisation purposes. In contrast to the work presented in [19] where the authors characterise short texts in order to identify possible drawbacks for the clustering of documents, they attempt to categorise documents within the same collection as opposed to our approach in which we focus on identifying the best features to determine if a set of documents is part of any of the categories already discussed. Our approach is based on lexical and statistical analysis which is beneficial to the process in order to be scaled.

There is no direct comparison between our work against other characterisation approaches due to the fact that other approaches apply their feature identification for different purposes such as topic identification in weblogs based on linkage analysis [12]. In [11] the authors characterise weblogs based on Web searching, Web pages and queries in order to identify topics and reading level distributions.

The purpose of characterisation has been adapted to particular tasks; in our case we provided features that showed to be useful for categorisation of short text collections.

As far as we are aware, there is no other work having the specific aim of identifying the best features of different short text genres in order to improve the categorisation task. Most of the related works are focused on a single genre of text. For this reason we have compared the three approaches so the results of using the best combination of features, discussed in this work, can be observed in Table 15.

6. Conclusion

Characterising the blogosphere is a highly challenging task and can be used to improve the quality of the categorisation task. The blogosphere in particular presents us with certain unique challenges. While this presents specific difficulties for classification and clustering algorithms, having a priori characterisation knowledge enables us to evaluate the collections and provide the categorisation process with the appropriate information for this process. It also allows us to take remedial action by choosing the most appropriate classification technique if it is needed.

We have focused our efforts on the extraction of features and their application in improving the accuracy of the categorisation process. We have covered a set of features that are important for characterising short texts. These features were grouped in five categories: Length-based, Vocabulary-based, POS based, Formality-based and Statistical-based. We present the results of the experiments in which we analyse the features of a sample weblog corpora, and compare the results to other types of short texts, extracted from more formal sources, such as newsfeeds and scientific publications. We used these characteristics in order to categorise different genres of short texts in a supervised manner. We also presented the performance of an unsupervised algorithm using the four sets of features. We were able to identify the important features which can be used to organise different genres of short texts in an automatic manner by using an unsupervised or a supervised approach. We tested the well-known Gaussian and SVM classifiers and the K-means clustering algorithm with several hundreds of different combinations of the features, in order to identify the best combinations of features.

After analysing the results of our experiments we may conclude that weblogs have very particular features in comparison with other types of short texts. The brevity of text, low vocabulary frequency, and the freedom in the writing style allowing grammar mistakes, misspellings and abbreviations are characteristics of weblog texts which make categorisation process difficult. Thus, selection of the appropriate features becomes even more important in order to feed the categorisation processes with the relevant information.

Even though the clustering approach did not outperform the classification approaches, we still believe that it is the best option for categorising short text collections because the difference between the results is not huge and the most important reason is that it is unsupervised approach. This fact can be very important in many scenarios such as narrow and specialised domains that we believe it is difficult to get label data for the training process of the classifier. In other

Table 15. Comparison of the best f-measure values.

	K-means	Features Used (K-means)	Gaussian Classifier (GC)	Features Used (GC)	SVM Classifier	Features Used (SVM)
Best Automatic Combination	0.63	LD, FT(all), TL (all), DL, iM	0.67	LD, TL (all), DL, F1, PPL, iM	0.66	LD, OOV, TL(all) DL, F1, iM
Empirical Selection	0.66	LD, FT(=1), TL (<4), DL, POS(NN, NNS, PRP, DT and ":"), iM	0.69	LD, FT(=1) TL (<4), DL, F1, PPL, POS(NN, NNS, PRP, DT and ":"), iM	0.70	LD, FT(=1), TL(<4), OOV, DL, F1, POS(NN, NNS, PRP, DT and ":"), iM

words, the benefit of using an unsupervised approach for categorising weblogs is evident based on the diversity of topics contained in these kinds of texts and the difficulty of providing external resources for training supervised approaches.

Acknowledgements

The research work of the third author is partially funded by the WIQ-EI (IRSES grant n. 269180) and DIANA APPLICATIONS (TIN2012-38603-C02-01), and done in the framework of the VLC/Campus Microcluster on Multimodal Interaction in Intelligent Systems.

References

[1] Amigo, E., Artiles, J., Gonzalo, J., Spina, D., & Liu, B., WePS-3 Evaluation Campaign: Overview of the Online Reputation Management Task. In 2nd Web People Search Evaluation Workshop (WePS 2010). Padova, Italy: CLEF 2010 Conference, 2010.

[2] A. Bell, Language Style as Audience Design. Language in Society, 13(2) (1984), 145–204.

[3] S. Brown, How Long is the Ideal Blog Post?, (2007), [Online] Available at: http://modernl.com/article/how-long-is-the-ideal-blog-post [Accessed June 2013].

[4] K. Burton, A. Java, and I. Soboroff, The ICWSM 2009 Spinn3r Dataset, In Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009), ACM, San Jose, CA, (2009).

[5] N. Christianini, and J. Shawe-Taylor, Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University, New York, 2000.

[6] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida, Traffic Characteristics and Communication Patterns in Blogosphere, International Conference on Weblogs and Social Media. AAAI, Colorado USA, (2007).

[7] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification, Wiley-Interscience, 2000.

[8] G. Grefenstette, Explorations in Automatic Thesaurus Discovery, Kluwer Academic Publishers Norwell, MA, USA, 1994.

[9] F. Heylighen and J. Dewaele, Formality of Language: definition, measurement and behavioral determinants. Free University of Brussels, Center "Leo Apostel", (1999).

[10] D. Ingaramo, and M. Errecalde, Departamento de informática. [Online], from Inteligencia Artificial, Available at http://www.dirinfo.unsl.edu.ar/~ia/resources/micro4news.rar [Accessed 2013].

[11] J. Y. Kim, K. Collins-Thompson, P. N. Bennett, and S. T. Dumais, Characterizing web content, user interests, and search behavior by reading level and topic, In Proceedings of the fifth ACM international conference on Web search and data mining, ACM, Seattle, Washington, USA, (2012).

[12] Z. Kostas, V. Vasiliki, and V. Dimitrios, Bloggers' Community Characteristics and Influence within Greek Political Blogosphere, Future Internet, Greece, (2012) 4, 396–412.

[13] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, RCV1: A New Benchmark Collection for Text Categorization Research, Journal of Machine Learning Research, (2004) 5, 361–397.

[14] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, Berkeley, University of California Press, (1967), 281–297.

[15] D. C. Manning, and H. Schütze, Foundations of statistical natural language processing, MIT Press MA, USA, 1999.

[16] A. Mosquera, and P. Moreda, The Use of Metrics for Measuring Informality Levels in Web 2.0 Texts, The 8th Brazilian Symposium in Information and Human Language Technology (STIL). Bazil: Sociedade Brasileira de Computacao, (2011).

[17] F. Perez-Tellez, D. Pinto, J. Cardiff, and P. Rosso, Characterizing Weblog Corpora, Lecture Notes in Computer Science, Springer, Germany, (2009), 299–300.

[18] F. Perez-Tellez, D. Pinto, J. Cardiff, and P. Rosso, Clustering Weblogs on the Basis of a Topic Detection Method, Springer-Verlag, Puebla, Mexico, (2010), 342–351.

[19] D. Pinto, P. Rosso, and H. Jimenez-Salasar, On the Assessment of Text Corpora, In Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems, NLDB-2009. LNCS (5723) Springer-Verlag, (2009), 281–290.

[20] Y. Qiu, and H. Frei, Concept based query expansion, In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieva, ACM, (1993), 160–169.

[21] A. Radford, M. Atkinson, D. Britain, H. Clahsen and A. Spencer, Word classes, In Linguistics An Introduction, Cambridge University Press, (2009), 125–139.

[22] C. Wartena and R. Brussee, Topic Detection by Clustering Keywords, In Proceedings of the 19th International Conference on Database and Expert Systems Application, IEEE Computer Society, USA, (2008), 54–58.

[23] D. Watson, Death Sentence: The Decay of Public Language, Random House Australia P/L, 2003.

[24] Y. Yang, 8 Websites To Shorten Your Tweets Automatically. [Online] Available at: http://freenuts.com/8-websites-to-shorten-your-tweets-automatically/ [Accessed June 2013].