Enhancing the discovery of informality levels in Web 2.0 texts

Alejandro Mosquera[*], **Paloma Moreda**[*]

[*]University of Alicante
University of Alicante, DLSI. Ap.de Correos 99. E-03080 Alicante, Spain
{amosquera, moreda}@dlsi.ua.es

Abstract

Social media publications are a popular and valuable source of information for Natural Language Processing applications. The linguistic analysis of these texts is a challenging task as a consequence of their informal nature. This paper explores the characterization of informality levels in Web 2.0 texts using unsupervised machine learning techniques. We prove that this task can be enhanced by introducing new text characteristics, like readability or emotional distance, and the evaluation of clustering quality indexes. The results of this approach are three informality levels and the characterization of their most remarkable features.

1. Introduction

In the social media the texts are public and open to comments for a large and unknown audience. In this context, the ability to share thoughts and ideas is more important than the formal content. Thus, the writing style of Web 2.0 can differ from traditional media, like newspapers or magazines, by being significantly more informal.

The study of text informality is relevant to a number of problems in Natural Language Processing (NLP). For discovering its impact on the state of the art NLP systems and tools we need a detailed knowledge of its main characteristics. However, there are few works directly related with this topic. For this reason, in this paper we are going to propose new methods for determining additional informality levels in Web 2.0 texts using an informality score and unsupervised machine learning techniques. In order to do this, we employ clustering quality indexes to determine the optimal number of levels and text features like readability measures or automatic emotion annotation to explore the obtained results.

This article is organized as follows: In Section 2 we review the state of the art. Section 3 describes our methodology. In Section 4, the obtained results are analysed. Finally, our main conclusions and future works are drawn in Section 5.

2. Related Work

Every speaker is normally in command of several different language registers. These will vary according to the topic under discussion, the field of communication and the formality context, being the term register commonly used as shorthand for formal/informal style. Hence, the definition of informality, and subsequently of formality, is directly related with the dimensions of language registers.

One of the most notable attempts to define language characteristics like registers is from Halliday, who determined the level of formality with three variables: field, tenor and mode (Halliday and Ghadessy, 1988).

Biber contributed important studies to the field of register, and the closely related genre, analysis (Biber, 1988). In his work, Biber makes use of Multidimensional Analysis (MDA), a complex methodology based on factor analysis, to study the dimensions of register variation for the English language, identifying 23 registers corresponding to the written and spoken English.

Other different studies about language registers rely on pattern analysis (Tribble, 1999), in which with the use of WordSmith (Scott, 1999) extracts keywords for grouping texts by the most relevant words, obtaining results similar to the MDA but without the factor analysis complexity.

More recent works are using machine learning techniques. In (Gries et al., 2009) an agglomerative hierarchical clustering of n-grams is used to differentiate registers for the English language.

Concerning Internet text types, previous Biber works were adapted, employing the MDA methodology in conjunction with cluster analysis for discovering language registers in Web texts (Biber and Kurjian, 2007).

While register analysis can give us information about the informality of a text, there are also measures that can be used for obtaining additional classifications. The most common formality metric is the word length, which has been used in other NLP tasks like genre classification (Karlgren and Cutting, 1994). The frequency of tokens present in formal and informal lists of words can also be used to determine word-level formality. In (Brooke et al., 2010) they evaluate different approaches for building formality lexicons using three basic features: word length, word count, and word association.

More complex measures usually involves formulae based on text features. One of the more remarkable scores is the F-Measure (Heylighen and Dewaele, 1999), using the concept of lexical density (Ure, 1971) with Part of Speech (POS) tags.

Readability metrics can be also used as a formality score. In (Lahiri et al., 2011) they show the existing correlation between formality and readability, experimenting with the sentence-level formality of texts from Web sources using the F-Measure and readability indexes.

These metrics can quantify how informal a text is but lack the register analysis granularity. Moreover, the language registers approaches discussed earlier, are usually complex and not informality-specific studies that focus on a higher number of dimensions.

For these reasons, we propose a method for obtaining different informality levels in Web 2.0 texts in order to discover their relevant characteristics. To achieve this, unsupervised machine learning techniques are used in conjunction with new metrics: an informality measure and cluster quality indexes.

3. Methodology

With the aim of enhancing the characterization of informality levels in social media texts, a set of text characteristics were developed and extracted from the dataset described in 4.2. Subsequently, we performed a dimensionality reduction of the obtained characteristics. After the repeated application of clustering processes on the only remaining dimension, we evaluated the optimal number of informality levels and analysed the final results.

This section is structured as follows: In 3.1 we describe the explored text characteristics. The informality measure used for clustering is introduced in section 3.2. Finally, the used classification algorithm is shown in section 3.3.

3.1. Text Characteristics

There are different studies about characterizing the features of formal or informal writing (Evans et al., 2004) (Thayer et al., 2010), they have in common the use of POS, word-length or sentence-length features. We defined a list of 21 interesting text characteristics to explore their variation in different informality levels:

(C1) Coleman-Liau: This index (Coleman and Liau, 1975) is a readability test designed to measure the understandability of a text. Its formula is defined as follows:

$CLI = 0,0588L - 0,296S - 15,8$

Where *L* is the average number of letters and *S* is the average number of sentences per 100 words.

(C2) RIX: This index measures text readability (Anderson, 1983) and is based on two factors, the length of words and the sentence length:

$RIX = LW/S$

Where *W* is the number of words, *S* is the number of sentences, and *LW* the number of words with more than 7 letters.

(C3) Emotional distance: On one hand, there are texts with neutral or low emotional loading, having a large distance with the reader. On the other hand, the texts that express feelings or emotions in direct or indirect manner are closer to the reader.

Using a similar approach such as (Park et al., 2011) where an automatic annotation of emotions in movie dialogues was performed, we developed a feature for measuring the text emotional distance using WordNet (Fellbaum, 1998) and a custom emotion lexicon. To achieve this, we computed the semantic similarity of each word (van Willegen et al., 2009) in the text respect to all the words contained in a lexicon of basic feelings. The lexicon was designed as a tree of seed words based on 6 primary emotions: *Love, Joy, Surprise, Anger, Sadness and Fear* (Parrott, 2001).

The semantic similarity for nouns is calculated using WordNet, obtaining the relatedness of each word respect the emotion list and keeping the score with the higher value. In case of adjectives, verbs and adverbs WordNet definitions are used to recursively calculate the score of the whole sentence and taking the higher value as the word similarity.

For words not present in WordNet a search in Roget's Thesaurus (American Psychological Association, 2011) is performed to obtain the word definition.

The score is averaged from all the words in the text to obtain the final emotional distance. A high value means low semantic distance to basic emotions thus closer to the reader, for the other side, a low value implies higher semantic distance to emotions thus larger distance with the reader.

(C4) Wrong-written sentences: We use 3 heuristic rules to determine ill-formed sentences (Lloret, 2011): Each sentence must contain at least 3 words, each sentence must contain at least 1 verb and the sentences cannot end in articles, prepositions or conjunctions.

(C5),(C6) Word length: Average words per sentence and average sentence length.

(C7),(C8),(C9) Part of speech: Frequency of grammatical part of speech tags obtained with TreeTagger (Schmid, 1994): Passive voice verbs, Action verbs [1] and Interjections.

(C10),(C11),(C12),(C13),(C14) Special words: Frequency of emoticons, informal, slang, offensive and unknown words classified by dictionary tags using Wiktionary [2], Online Slang Dictionary [3] and Advanced Learner Cambridge Online Dictionary [4].

(C15) Wrong-typed words: We use simple heuristics for detecting wrong-typed words taking into account their case and position in the sentence (*YoU, ARE sure? no... i'm not*)

(C16) Frequency of stopwords

(C17) Frequency of contractions: (*can't, It's...*).

(C18) Frequency of formal words [5]

(C19) Unknown words: Frequency of words marked by the POS-Tagger lemmatizer as unknown.

(C20) First and second pronouns: Frequency of first and second person personal pronouns.

[1] http://rfptemplates.technologyevaluation.com/List-of-Action-Verbs.csv

[2] http://en.wiktionary.org

[3] http://onlineslangdictionary.com

[4] http://dictionary.cambridge.org

[5] http://www.plainlanguage.gov/howto/wordsuggestions/simplewords.cfm

(C21) F-Measure: The formality measure F-Measure defined by Heylighen as follows:

F-Measure = (noun frequency + adjective freq. + preposition freq. + article freq. - pronoun freq. - verb freq. - adverb freq. - interjection freq. + 100)/2

3.2. Feature Reduction

We use the informality score I-Measure (Mosquera and Moreda, 2011b), a variable obtained with factor analysis (Rummel, 1970) and Principal Component Analysis (PCA) (Jolliffe, 2002) to reduce the dimensionality of our text characteristics. The resultant feature represents the most of the variance and avoids correlation for obtaining the best results in the clustering process.

*I-Measure = (Wrong-typed Words freq. + Interjections freq. + Emoticon freq.) * 100*

3.3. Classification Algorithm

The Expectation-Maximization (EM) (Dempster et al., 1977) algorithm was used to cluster our reduced dataset containing the I-Measure as the unique feature. The EM algorithm is an iterative optimization method that estimates missing parameters of probabilistic models. This is generally a two step optimization approach, performing an initial approximation of the cluster parameters:

- The expectation step (E-Step) calculates the expected value $Q(\theta, \theta^k)$ of the log like-hood function $\log p(y, z|\theta)$.
- The maximization step (M-Step) calculates the distribution parameters θ^{k+1} and their likelihood $Q(\theta, \theta^k)$.

4. Evaluation and Results

For conducting the evaluation process, in the next subsection (4.1) we introduce the used evaluation measures. Section 4.2 describes the used corpora. The current baseline is explained in section 4.3. The experimental results are analysed in section 4.4 and interpreted in section 4.5.

4.1. Evaluation Measures

Assessing the validity of a number of clusters relative to others is an important issue in cluster analysis. There exists two basic approaches: Internal criterion favours clusters with high intra-cluster similarity and low inter-cluster similarity, external criterion evaluate performance against an external benchmark. In this particular case, we chose 3 internal indexes for evaluating clustering quality, each one measuring different clustering features and avoiding the need of costly manual annotations.

The **Silhouette validation** technique (Rousseeuw, 1987) is a graphical representation of tightness and separation of clusters in a data set. To calculate the silhouettes *S(i)* we use the following formula:

$$S(i) = \frac{(b(i) - a(i))}{max(a(i), b(i)} \qquad (1)$$

Where *a(i)* is the average dissimilarity of *i* with all other data within the same cluster and *b(i)* is the minimum of the average dissimilarity of *i* to all objects in other cluster.

The average silhouette width $(-1 \leq S(i) \leq 1)$ can be used to evaluate the clustering results and determine the most appropriate set of clusters. If the silhouette width is close to -1, it means that the sample is not well classified. If the silhouette width is close to 1, the sample was assigned to a very appropriate cluster.

The **Dunn index** (Dunn, 1974) measures isolation and compactness in cluster sets. To calculate the Dunn validation index, *D*, we use the following formula:

$$D = \min_{1 \leq i \leq n} \{ \min_{1 \leq j \leq n, i \neq j} \{ \frac{d(c_i, c_j)}{\max_{1 \leq k \leq n}(d'(c_k))} \} \} \qquad (2)$$

Where c_i represents the cluster of the i-partition, $d(c_i, c_j)$ is the distance between clusters c_i and c_j, $d'(c_k)$ is the intra-cluster distance of cluster c_k and n is the number of clusters. A good clustering maximizes the inter-cluster distances and minimise the intra-cluster distances. Therefore, the number of clusters that maximize D is taken as the optimal number of the clusters.

The **Connectivity measure** (Handl et al., 2005), uses the concept of neighbourhood for measuring the amount of *connectedness* of a particular cluster. Using this criteria means that neighbouring data items should share the same cluster. The Connectivity index can be calculated with the following formula:

$$Conn(\varrho) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i, nn_{i(j)}} \qquad (3)$$

Where *L* is the number of nearest neighbours to use, *N* is the number of objects, $nn_{i(j)}$ is the jth nearest neighbour of observation *i* and $\varrho = \{C_1 .... C_k\}$ a particular clustering of the *N* observations into *K* disjoint clusters. This index has a value between $(0, \infty)$ considering the clustering that minimizes the connectivity as the optimal solution.

4.2. Corpus Characteristics

A subset of the CAW2.0 dataset (Fundacion Barcelona Media, 2009) was used to obtain 7000 texts from the following Web 2.0 sources: **Slashdot**, a technology-related news website; **Ciao**, an on-line shopping and product review portal; **Kongregate**, an on-line gaming and chat website; **Twitter**, a social networking and microblogging service; **MySpace**, a social networking website; **Digg**, a news voting and review website; and **Engadget**, and electronic products review portal. In addition we included 1000 texts from news comments of **TheGuardian**, an on-line newspaper.

Also we used the first 100 texts of the Penn Treebank corpus for English language (Marcus et al., 1994) for evaluating our results against a non-Internet and moderately formal collection of texts.

4.3. Baseline

Our current baseline is built upon our previous studies (Mosquera and Moreda, 2011a) and (Mosquera and

Nº Clust.	Silh.	Dunn	Connectivity
2	0,29	0.00004638656	3.857937
3	**0,71**	**0.002040972**	**3.503571**
4	0,61	0.00023	11.86310
5	0,66	0.00008527276	16.87897
6	0,65	0.00005564708	23.26944
7	0,62	0.00002647555	25.09921
8	0,63	0.00004374222	21.64484
9	0,67	0.0008611206	24.37540
10	0,66	0.0002201764	35.16548

Table 1: Cluster validation results.

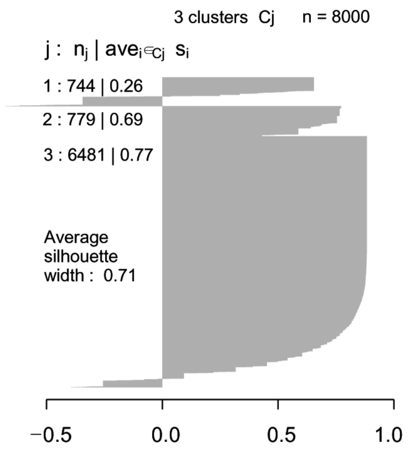

Figure 1: Silhouette width validation results for 3 clusters.

Moreda, 2011b), where in our first approaches we established two informality levels using smaller datasets (350 and 700 texts), a more simple set of text features (POS, special words, word and sentence length, emoticons) and unsupervised machine learning algorithms.

4.4. Results

The dataset was partitioned with the EM algorithm from 2 to 10 clusters, then we calculated the clustering quality indexes for each cluster (see Table 1). The average silhouette width (see Figure 1) shows a well formed clustering using 3 partitions (0,71 Avg. SW), specially in clusters 2 and 3, being weaker in cluster 1. Regarding Dunn and Connectivity indexes they also obtain their best values with the same number of partitions. So we can conclude that the best clustering result is obtained with 3 clusters, improving our two-cluster baseline.

The distribution of the average clustered characteristics can be compared with the average results of our subset of Penn Treebank English Corpus (see Table 2) to understand better the differences between the 3 informality levels in Web 2.0 texts and non-Web texts.

4.5. Interpretation

After analysing the obtaining results we can extract some direct conclusions (see Table 3), obtaining three informality levels: The first level can be still considered as informal in comparison with a non-Internet corpus, but with slightly formal characteristics like the use of passive verb voice and semi-elaborated sentences (high frequency of stopwords and the presence of formal words). The second one is an intermediate level, formal and informal features appear mixed but with prevalence of informal content (frequency of contractions and the use of the first and second person pronouns). The third and last level congregates the more informal and low quality texts, with high presence of slang, offensive and unknown words, very short words and sentences, abundance of typos, wrong-constructed sentences and very short distance with the reader.

Nº	Cluster3	Cluster1	Cluster2	Penn
(C1)	5,672	-0,403	-8,431	10,634
(C2)	2,501	0,561	0,379	6,186
(C3)	0,033	0,068	0,090	0,005
(C4)	0,058	0,130	0,182	0,010
(C5)	15,755	4,958	2,439	23,834
(C6)	64,503	18,496	9,894	110,404
(C7)	0,001	0,000	0,000	0,005
(C8)	0,016	0,010	0,002	0,011
(C9)	0,006	0,047	0,109	0,000
(C10)	0,001	0,009	0,019	0,000
(C11)	0,004	0,012	0,038	0,002
(C12)	0,022	0,029	0,045	0,038
(C13)	2,007	4,108	16,622	0,000
(C14)	0,048	0,083	0,258	0,052
(C15)	0,024	0,212	0,721	0,008
(C15)	0,403	0,283	0,092	0,285
(C17)	0,004	0,020	0,008	0,002
(C18)	0,006	0,003	0,002	0,022
(C19)	0,079	0,179	0,483	0,000
(C20)	0,031	0,041	0,007	0,000
(C21)	58,959	50,548	50,633	121,899
(I*)	3,1	26,8	84,8	0,8

Table 2: Averaged values for the clustered data and the Penn Treebank corpus. (* I-Measure)

5. Conclusions and Future Works

The results obtained in this paper show interesting characteristics that would be difficult to obtain with traditional approaches for generic text types. We discovered remarkable differences between the 3 informality levels like readability, emotional distance or frequency of typos that must be taken into account when processing Web 2.0 texts. Moreover, the development of new features like emotional distance has been proved useful, extracting relevant semantic information from very low quality texts. In addition, the evaluation of internal clustering measures avoids the need of costly and error-prone tasks like manual annotations, allowing the use of large corpora.

Future works about this topic would include the performance measurement of NLP systems and applications in the 3 obtained informality levels and an improved emotional analysis. Moreover, a fine-grained study of each text type is the next step for obtaining a more detailed and complete classification.

Acknowledgements

This paper has been partially supported by Ministerio de Ciencia e Innovación - Spanish Government (grant no. TIN2009-13391-C04-01), and Conselleria d'Educación - Generalitat Valenciana (grant no. PROMETEO/2009/119, ACOMP/2010/286 and ACOMP/2011/001)

6. References

American Psychological Association, APA, 2011. Roget's 21st century thesaurus, third edition.

Slightly Informal (Cluster3)	Moderately Informal (Cluster1)	Very Informal (Cluster2)
Mixes long and short sentences	Predominance of short sentences	Only short sentences
Use of formal words	Low use of formal words	Very low use of formal words
High presence of stop-words	Presence of stop-words	Low presence of stop-words
Use of passive voice verbs	Absence of passive voice verbs	Absence of passive voice verbs
Easy to read	Very easy to read	Texts are extremely simple
Use of first and second pronouns	Use of first and second pronouns	Low use of first and second pronouns
Low use of contracted forms	Frequent use of contracted forms	Use of contracted forms
Presence of unknow words	High presence of unknow words	Very high presence of unknow words
Short distance with the reader	Very short distance with the reader	No distance with the reader
Use of slang and offensive words	High use of slang and offensive words	Very high use of slang and offensive words
Presence of typos	High presence of typos	Very high presence of typos
Use of interjections and emoticons	Moderate use of interjections and emoticons	High use of interjections and emoticons

Table 3: Features of the the three obtained informality levels

Anderson, J., 1983. Lix and rix: variations on a little-known readability index. *Journal of Reading*, 26(6):490–497.

Biber, D., 1988. *Linguistic features: algorithms and functions in Variation across speech and writing*. Cambridge University Press.

Biber, D. and J. Kurjian, 2007. Towards a taxonomy of web registers and text types: A multi-dimensional analysis. In In M. Hundt N. Nesselhauf and C. Biewer (eds.), *Corpus linguistics and the web*. Amsterdam, Rodopi, pages 109–132.

Brooke, Julian, Tong Wang, and Graeme Hirst, 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics.

Coleman, M. and T. L. Liau, 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Dempster, A. P., M. N. Laird, and D. B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–22.

Dunn, J. C., 1974. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57.

Evans, M.B., A.A. McBride, M. Queen, A. Thayer, and J.H. Spyridakis, 2004. The effect of style of typography on perceptions of document tone. *Proceedings of IEEE International Professional Communication Conference*:300–303.

Fellbaum, Christiane (ed.), 1998. *WordNet An Electronic Lexical Database*. Cambridge, MA ; London: The MIT Press.

Fundacion Barcelona Media, (FBM), 2009. Caw 2.0 training datasets.

Gries, Stefan Th., John Newman, and Cyrus Shaoul, 2009. N-grams and the clustering of registers. *ELR Journal*, 5.

Halliday, M.A.K. and Mohsen Ghadessy, 1988. *On the language of physical science. In Mohsen Ghadessy (ed.), Registers of Written English: situational factors and linguistic features*. London and New York: Pinter Publishers. 162-178.

Handl, Julia, Joshua D. Knowles, and Douglas B. Kell, 2005. Computational cluster validation in post-genomic data analysis. *Bioinformatics*:3201–3212.

Heylighen, Francis and Jean-Marc Dewaele, 1999. Formality of language: definition, measurement and behavioral determinants. Technical report, Free University of Brussels.

Jolliffe, I. T., 2002. *Principal Component Analysis*. Springer, 2nd edition.

Karlgren, Jussi and Douglas R. Cutting, 1994. Recognizing text genres with simple metrics using discriminant analysis. In *COLING'94*.

Lahiri, Shibamouli, Prasenjit Mitra, and Xiaofei Lu, 2011. Informality judgment at sentence level and experiments with formality score. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*, CICLing'11. Springer-Verlag.

Lloret, Elena, 2011. Text summarisation based on human language technologies and its applications. *Ph.D. dissertation. University of Alicante*.

Marcus, Mitchell P., Beatrice Santorini, and Mary A. Marcinkiewicz, 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19.

Mosquera, Alejandro and Paloma Moreda, 2011a. Caracterización de niveles de informalidad en textos de la web 2.0. *Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, 47.

Mosquera, Alejandro and Paloma Moreda, 2011b. The use of metrics for measuring informality levels in web 2.0 texts. *Proceedings of 8th Brazilian Symposium in Information and Human Language Technology (STIL)*.

Park, Seung-Bo, Eunsoon Yoo, Hyunsik Kim, and GeunSik Jo, 2011. Automatic emotion annotation of movie dialogue using wordnet. In *ACIIDS (2)*.

Parrott, W., 2001. Emotions in social psychology.

Rousseeuw, Peter, 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65.

Rummel, R.J., 1970. *Applied Factor Analysis*. Evanston: Northwestern University Press.

Schmid, Helmut, 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Scott, M., 1999. Wordsmith tools version 3.

Thayer, Alexander, Mary B. Evans, Alicia A. McBride, Matt Queen, and Jan H. Spyridakis, 2010. I, pronoun: A study of formality in online content. *Journal of Technical Writing and Communication*, 40:447–458.

Tribble, Christopher, 1999. Writing difficult texts. *Ph.D. dissertation. Lancaster University*.

Ure, J., 1971. Lexical density and register differentiation. in g. perren and j.l.m. trim (eds), applications of linguistics:443–452.

van Willegen, Ivar, Léon J. M. Rothkrantz, and Pascal Wiggers, 2009. Lexical affinity measure between words. In *TSD*.