AAAI Technical Report WS-12-02
Real-Time Analysis and Mining of Social Streams

SMILE: An Informality Classification Tool for Helping to Assess Quality and Credibility in Web 2.0 Texts

Alejandro Mosquera
University of Alicante
Alicante, Spain
amosquera@dlsi.ua.es

Paloma Moreda
University of Alicante
Alicante, Spain
moreda@dlsi.ua.es

Abstract

The data made available by Web 2.0 applications such as social networks, on-line chats or blogs have give access to multiples sources of information. Due to this dramatic increase in available information, the perception of quality and credibility plays an important role in social media, thus making necessary to discard low quality and uninteresting content. Moreover, the informal features of Web 2.0 texts such as emoticons, typos, slang or loss of formatting impact negatively on user perception regarding content quality and credibility. For this reason, this paper proposes the SMILE system, a novel unsupervised real-time tool for assessing user-generated content quality and credibility using informality levels. As a test case, we focus on Yahoo! Answers, a relevant Web 2.0 application by its amount of users, content and textual diversity. The results of our study show that informality analysis can be used as criteria to help assess the credibility and quality of Web 2.0 information sources.

1 Introduction

User-generated content transformed the way that information is handled in Internet. This paradigm shift focused in the user has given rise to the Web 2.0, where users generate, share and consume information. The data made available by Web 2.0 applications such as social networks, on-line chats or blogs has give access to multiples sources of information.

Due to this dramatic increase in available information, the perception of quality and credibility plays an important role in social media, thus making necessary to discard low quality and uninteresting content.

In order to evaluate content quality and credibility, three main elements can be identified: i) Content visibility or prominence is related to likelihood that a Web site element will be noticed or perceived (Fogg 2003); ii) Author expertise and trustworthiness can be used to measure credibility (Hovland, Irving, and Harold 1953). While trustworthiness can be more subjective, expertise includes message characteristics such as information quality or the use of titles that certify the communicator skills and experience on the topic (Mattus 2007); iii) Interpretation is a factor that relies on user judgement about online content (Fogg 2003). Because its relevance for user-generated content, this will be the factor which this study will focus on.

It is common to find in Web 2.0 applications the use of ranking metrics such as positive/negative votes, "likes" or number of shares in order to allow users to share feedback and opinions about user-generated content.

The non-standard characteristics present in these texts such as: lack of punctuation, loss of formatting, use of slang, colloquialisms, typos or emoticons (Baron 2003), remark the informal nature of the Web 2.0.

We propose that the informal features of texts, such as the present in Web 2.0 publications, impact negatively on content quality and credibility. For this reason, this paper presents the SMILE system (Social Media Informality Level Explorer) [1], a novel real-time tool for classifying Web 2.0 texts by their informality level.

This approach is based on unsupervised machine learning techniques and, as a test case, focused on Yahoo! Answers[2] publications. Yahoo! Answers is a question/answering site where people publicly post questions and answers on any topic. This site has more than 120 million users worldwide, has compiled 400 million answers and is the second-most-visited education/reference site on the Internet after Wikipedia (Leibenluft 2007).

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

2 Related Work

In this section we are going to talk about the main works related to text informality and credibility. On the one hand, the proposals for assessing on-line content quality and credibility will be detailed. On the other hand, informality analysis systems for Web 2.0 texts will be reviewed.

The use of informal text features such as emoticons, exclamations and capital letters has proven useful for credibility classification on Twitter texts (Castillo, Mendoza, and Poblete 2011). In addition, non-textual quality features such as votes, recommendations or click counts have been

[1] http://intime.dlsi.ua.es:8080/Smile/pages

[2] http://answers.yahoo.com/

used to determine page quality for Question and Answering (Q&A) websites (Jeon et al. 2006). Moreover, focusing only on metadata or genre-specific features do not take into account the special characteristics of the language present in user-generated content rather than generic Web 2.0 language characteristics.

Regarding the subject of informality analysis, we found little in the literature but a minimum amount in the opposite field study of text formality. A distinction between two different formality levels (formal/informal) (Lahiri, Mitra, and Lu 2011) can be performed using the F-Measure[3], one of the first scores for measuring text formality (Heylighen and Dewaele 1999). Another studies based on a two-level classification use machine learning (Sheikha and Inkpen 2010) or the concept of social distance and politeness (Peterson, Hohensee, and Xia 2011). Finally, there are works based on a word-level analysis of formality using lexicon dictionary methods (Brooke, Wang, and Hirst 2010). All these approaches tackle the subject from a formality point of view. Moreover, as user-generated content is mostly informal, an analysis of Web 2.0 non-standard language from an informality point of view is more appropriate (Mosquera and Moreda 2011).

For these reasons, in this article we propose the SMILE system, a novel unsupervised and real-time tool for classifying Web 2.0 texts by their informality level for helping to assess content quality and credibility.

3 SMILE Description

SMILE is an informality classification tool, and its main purpose is to help assess quality and credibility in Web 2.0 texts. To our knowledge, SMILE is the first tool designed to study the informality levels of textual content in social media. In this particular case we are going to perform an informality analysis of Yahoo! Answers texts, but the proposed system is not exclusive of this Web 2.0 application. The system presented in this work, named SMILE, makes use of unsupervised machine learning techniques to automatically classify Yahoo! Answers by their informality level.

In section 3.1 we introduce the Yahoo! Answers application. In section 3.2 we explain the classification process.

3.1 Yahoo! Answers

Yahoo! Answers is a user-moderated site where people ask and answer questions on any topic. Once a question has been answered it can be marked as “best answer” and then the question cycle is closed. However, answers can also be rated by the community to become the best answer and marked using a “thumbs up” or “thumbs down” with a positive or negative vote. Users can also award a question with stars, as a flag of being an interesting question. The system rewards users with points for posting and evaluating questions or answers. This scoring system serves as motivation tool and also as a credibility measure. The most voted answers and the answers of users with higher number of points tend to be considered more reliable.

A subset of 14500 texts extracted from the L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0 dataset [4] has been used. It includes all the questions and their corresponding answers. This corpus contains a small amount of metadata: Which answer was selected as the best answer, and the category and subcategory which was assigned to this question.

3.2 Classification

The SMILE classification system uses three basic processes: Feature extraction, clustering and data retrieval.

Feature Extraction SMILE extracts a set of 12 text features from each answer such as:

RIX: This index measures text readability (Anderson 1983): $RIX = LW/S$

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

Entropy: Making use of information theory, texts with high entropy would imply a higher amount of expressed information and more quality, otherwise texts with low entropy would communicate less information and can be considered of lower quality. This feature calculates de Shannon entropy (Shannon 1951) of each text.

Emotional distance: Using a similar approach as in (Park et al. 2011), we measure the emotion with its corresponding strength calculating the path distance of each word with WordNet (Fellbaum 1998) and an emotion lexicon based on 6 primary emotions: *Love, Joy, Surprise, Anger, Sadness and Fear* (Parrott 2001).

Incorrectly-written sentences: Heuristic rules to detect ill-formed sentences: Each sentence must contain at least three words, each sentence must contain at least one verb and the sentences can not end in articles, prepositions or conjunctions (Lloret 2011).

Misspelled words: We use simple heuristics for detecting common misspelled words taking into account their case and position in the sentence.

Frequency of contractions

Frequency of repetitions: Repetition of vocal or consonants within a word to add expressive power is used in informal speech and writing (*yeahhh, noooo!*).

Frequency of emoticons

Frequency of interjections: We use TreeTagger (Schmid 1994) to extract this part of speech.

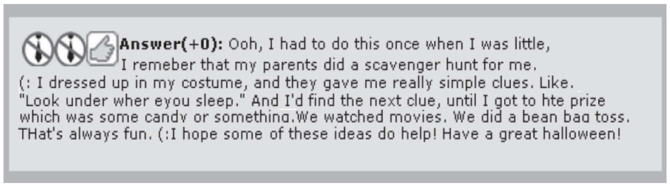
Frequency of upper-case words

Frequency of slang and offensive words: They are detected by a custom lexicon created with entries from [5] an on-line slang dictionary and a list of common offensive words.

[3] Do not confuse with the well-known F-measure in the field of Information Retrieval.

[4] http://webscope.sandbox.yahoo.com/

[5] http://onlineslangdictionary.com



Answer(+0): Ooh, I had to do this once when I was little,
I remeber that my parents did a scavenger hunt for me.
(: I dressed up in my costume, and they gave me really simple clues. Like.
"Look under wher eyou sleep." And I'd find the next clue, until I got to hte prize
which was some candy or something.We watched movies. We did a bean bag toss.
THat's always fun. (:I hope some of these ideas do help! Have a great halloween!

Figure 1: Example of SMILE answer classification. The number of crossed tie icons represents the text informality level. A green thumb-up icon is present when the answer is marked as the best by the community.

Clustering The application of unsupervised machine learning has the advantage of not depending on manually-annotated corpora. For this reason, we use the Expectation-Maximisation (EM) (Dempster, Laird, and Rubin 1977) unsupervised machine learning algorithm. With the corpus explained in 3.1 and extracting the features described in 3.2, a four-cluster model was obtained by using v-Fold cross validation (Witten and Frank 2005) and the implementation of EM provided by Weka (Hall et al. 2009).

In order to determine the optimal number of clusters the dataset was partitioned into 10 equally sized folds (9 for training and 1 for testing). Starting the number of clusters *k* with an initial value of 1, the EM clustering was performed for each set of 9 folds and the log-likelihood of the test sets was calculated. If the average log-likelihood increases from the previous value, *k* is then incremented by 1 and the process repeated.

Finally, each question analysed by the SMILE system is evaluated against our four-cluster model, mapping the assigned cluster to its informality level (see Table 1).

Cluster	Informality Level
Cluster 1	Level 1 (Low informality)
Cluster 2	Level 2 (Slight informality)
Cluster 3	Level 3 (High informality)
Cluster 4	Level 4 (Very high informality)

Table 1: Clusters and informality levels.

Data retrieval When an user introduces a question in SMILE a real-time query is performed to Yahoo! Answers on-line database using Yahoo Query Language (YQL) [6] and all the answers related to the matched questions are retrieved.

Because a tie can be a symbol of formality, we chose to represent the concept of informality with a crossed tie icon, thus the absence of formality. As we use a four-level classification scale, SMILE shows from one to four crossed ties, one for each informality level. In addition, we display the answer marked as best by the community with a green thumb-up icon and the average score for each answer (see Figure 1).

4 Evaluation and Results

We performed two types of analysis, in the first one we rely on the judgements of individual analysts, matching the informality classification results of our tool against manual annotations in order to obtain the precision and recall. In the second one, non-textual metadata features such as votes and "likes" rated by users are analysed with aim to explore their relation with the obtained informality levels.

4.1 Datasets

In order to evaluate SMILE results, two datasets have been used. The dataset A contains 50 answers selected by querying the top ten Yahoo! Search items [7] into Yahoo! Answers (see Table 2).

N.	Search item
1	Facebook
2	YouTube
3	NFL
4	Ashton Kutcher
5	Will Ferrell
6	Angelina Jolie
7	Jessica Simpson
8	NASCAR
9	NFL Scores
10	Peter Frampton

Table 2: List of popular items in Yahoo! Search used to generate the dataset A.

The dataset B contains 50 different answers selected by querying the top five Yahoo! Clues (Beta) items [8] into Yahoo! Answers (see Table 3).

N.	Search item
1	kim kardashian
2	selena gomez
3	miley cyrus
4	emma watson
5	jennifer lopez

Table 3: List of popular Yahoo! Clues (Beta) items used to generate the dataset B.

4.2 Inter-coder Agreement

A group of six people was asked to rate informality levels from more informal to less informal for the dataset A. We adopted a 1-4 Likert scale (Likert 1932), where 1 represents a low informal text, 2 a medium informal text, 3 a high informal text and 4 a very high informal text (see Table 4).

We evaluated the inter-coder agreement between the six raters in order to determine the degree of agreement (see Table 5). With a concern of not biasing the opinion of

[6] http://developer.yahoo.com/yql/

[7] http://buzzlog.yahoo.com/overall/

[8] http://clues.yahoo.com/

Answers	Level
Your link doesn’t work.So I cant see what it looks like,sorry. :]	2
dewd i would answer but i can’t u see ur link is bad.........	4
I don’t like herbut I guess she’s pretty and talented and all that stuff...	1
she owns a house???	2
Best Answer PLEASE I THINK I’M GONNA DIE PLEASE	3

Table 4: Examples of informality level classification with Yahoo! Answers.

the group, no additional classification information was provided. The use of nominal measures such as Cohen’s Kappa (Cohen 1960) or Fleiss’ Kappa (Davies and Fleiss 1982) shows no-agreement at all, and even taking advantage of the Likert scale, the ordinal Krippendorff’s (Krippendorf 2004) shows a low agreement value. These results remark the difficulty of analysing a complicated phenomenon, prone to different subjective analysis.

Measure	Overall
Fleiss’ κ	0.109
Average Pairwise Cohen’s κ	0.125
Average Pairwise Agreement	**0.340**
Krippendorff’s α	0.318

Table 5: Inter-coder agreement values.

As the previous assessment experiment was complicated due to the lack of a tagging scheme, which resulted in a high disagreement between raters, we asked to a different group of six people to rate the dataset B. In this second evaluation we provided the classification results obtained with SMILE, and raters were asked to confirm if they were agree with SMILE classification. In case of disagreement with the proposed informality levels, they should provide their own values. The results of this second inter-coder agreement evaluation showed an important increase even on nominal measures, reaching a 0.804 Krippendorff’s Alpha, a good agreement value (see Table 6).

Measure	Overall
Fleiss’ κ	0.522
Average Pairwise Cohen’s κ	0.524
Average Pairwise Agreement	0.640
Krippendorff’s α	**0.804**

Table 6: Inter-coder agreement values in the second evaluation.

4.3 SMILE Evaluation

The scores provided by the participants for the dataset B were matched with the informality levels provided by the developed tool (see Table 7). The best classifications results in terms of precision were obtained at the fourth informality level (82%), showing that SMILE can be effective in detecting high informality texts. Moreover, taking into account the overall results, SMILE classification scored a 60.6% in F1. These results can be considered positive taking into account the difficulty of the task and the initial low agreement between the different annotators.

Class	Precision	Recall	F1
Level 1	0.714	0.526	0.606
Level 2	0.722	**0.619**	**0.667**
Level 3	0.714	0.312	0.435
Level 4	**0.818**	0.450	0.581
Overall	0.740	0.513	0.606

Table 7: smile Informality level evaluation.

4.4 Metadata Evaluation

The use of non-textual metadata features such as votes, “likes” and user feedback can help the quality and credibility assessment of texts in question answering systems such as Yahoo! Answers (Agichtein et al. 2008). In order to analyse user perception about answer quality and credibility, two Yahoo! Answers variables have been selected in this study: the best answer flag and the number of “likes” or positive votes.

Analysing answers votes, answers marked as best by the community can be considered less informal than regular ones. The differences are significant taking into account the first and the last informality levels, where the 62% of the best answers were classified into the less informal level, while only a 45% of regular answers were classified into this level. Taking into account the more informal levels, the difference is even higher, with a 5% of best answers against a 10% of regular ones. These results are shown in Figure 2.

Regarding user “likes”, the correlation between informality and positive ratings is not always direct. While the answers classified into the first two informality levels are highly rated, a considerable number of positive ratings can be appreciated in the most informal texts (see Table 8). Analysing these results, we have found that controversial answers, jokes and even disqualifying comments are usually rated positive.

It could be concluded that there is a relation between Yahoo! Answers metadata and the informal language of its users. Moreover, the two evaluated application-specific variables for measuring credibility and quality obtained different results. While the use of votes has a direct relation with the informality levels, the feedback received from “likes” shows that this variable is not always employed by Yahoo! Answers

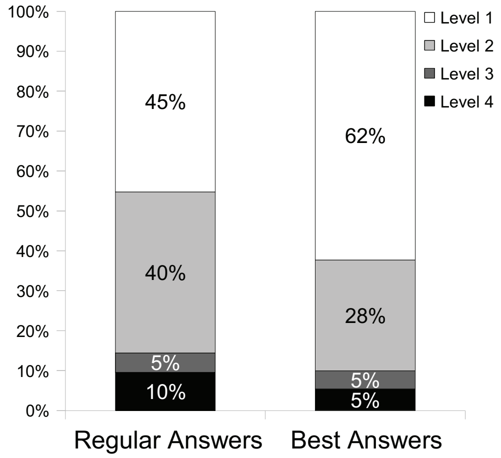


Figure 2: Distribution of answers by informality level

Informality Level	Average Rating
Level 1	1.749
Level 2	1.040
Level 3	0.277
Level 4	0.875

Table 8: Average answer rating by informality level.

users as a quality or credibility measure thus not following a linguistic criteria.

5 Final Remarks and Future work

In this paper, we have presented SMILE, a real-time tool for assessing user-generated content quality and credibility using informality levels. The evaluation of this tool has been performed on a real application using Yahoo! Answers. The obtained overall classification results (60.6% F1) showed that SMILE is suitable for this difficult task, taking into account that such classification is even challenging for manual reviewers.

It has been also shown that credibility is related with the informality phenomenon by analysing Yahoo! Answers metadata. The study of two credibility and quality variables such as the "best answer" flag and the number of "likes" confirmed our initial hypothesis: the informal features of texts impact negatively on user perception regarding content quality and credibility. It can be concluded that the obtained four-level informality classification can provide valuable information for the credibility and quality assessment of user-generated content in Web 2.0 applications.

In a future, we plan to develop an informality annotation guide with aim to improve both agreement and classification results. Also, the analysis of non-textual variables such as informal fonts, text colors or irregular formatting that can provide additional information is left to a future work.

Acknowledgments

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

References

Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008. Finding high-quality content in social media with an application to community-based question answering. In *In Proceedings of WSDM*.

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

Baron, N. S. 2003. Language of the internet. *The Stanford Handbook for Language Engineers, pp. 59-127*.

Brooke, J.; Wang, T.; and Hirst, G. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, 90–98. Stroudsburg, PA, USA: Association for Computational Linguistics.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. *Proceedings of World Wide Web Conference (WWW)*.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20(1):37.

Davies, M., and Fleiss, J. L. 1982. Measuring agreement for multinomial data. *Biometrics, 38(4)* 1047–1051.

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

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

Fogg, B. J. 2003. Prominence-interpretation theory: Explaining how people assess credibility online. *CHI '03 extended abstracts on Human factors in computing systems*.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.* 11:10–18.

Heylighen, F., and Dewaele, J.-M. 1999. Formality of language: definition, measurement and behavioral determinants. Technical report, Free University of Brussels.

Hovland, C. I.; Irving, J. L.; and Harold, K. H. 1953. *Communication and Persuation: Psychological Studies of Opinion Change*. New Haven: Yale UP.

Jeon, J.; Croft, W. B.; Lee, J. H.; and Park, S. 2006. A framework to predict the quality of answers with non-textual features. In *In SIGIR ?06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 228–235. ACM Press.

Krippendorf, K. 2004. *Content Analysis: An Introduction to Its Methodology*. Second edition, chapter 11. Sage, Thousand Oaks,CA.

Lahiri, S.; Mitra, P.; and Lu, X. 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, 446–457. Springer-Verlag.

Leibenluft, J. 2007. A librarian's worst nightmare: Yahoo! answers, where 120 million users can be wrong. *Slate Magazine*.

Likert, R. 1932. A technique for the measurement of attitudes. *Archives of Psychology 140* 1:55.

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

Mattus, M. 2007. Finding credible information: A challenge to students writing academic essays. *Human IT, 9(2)* 1–28.

Mosquera, A., and Moreda, P. 2011. 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, S.-B.; Yoo, E.; Kim, H.; and Jo, G. 2011. Automatic emotion annotation of movie dialogue using wordnet. In *ACIIDS (2)*, 130–139.

Parrott, W. 2001. *Emotions in Social Psychology*. Psychology Press, Philadelphia.

Peterson, K.; Hohensee, M.; and Xia, F. 2011. Email formality in the workplace: A case study on the enron corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 86–95. Portland, Oregon: Association for Computational Linguistics.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 44–49.

Shannon, C. E. 1951. *Prediction and entropy of printed English.* Number 30. The Bell System Technical Journal.

Sheikha, F. A., and Inkpen, D. 2010. Automatic classification of documents by formality. *Proceedings of the 2010 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE2010)* 1–5.

Witten, I. H., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Sys. Morgan Kaufmann, second edition.