Automating Affect Detection in Multilingual Text Chat

ABSTRACT

Investigating emotion in chat-based collaborative work is necessary for understanding team dynamics and interactions. Manual annotation can be scaled to larger datasets with experimental machine learning techniques. Recent work has made progress in increasing accuracy even in difficult contexts, like chat rooms, where utterances are short, and slang, abbreviations, misspellings, and lack of grammar are common. However, the increasing prevalence of geographically distributed, multicultural, and multilingual teams presents new challenges to automating affect detection and studying group dynamics based on chat room communication. In this paper, we introduce an adaptation of the statistical affect detection in collaborative chat approach, ALOE, presented at CSCW2013 [2]. This adaptation is intended to help detect affect in French chat while maintaining performance on English. We contribute a catalogue of modifications for using ALOE with French chat logs, as well as a discussion of the implications of automating affect detection across linguistic and cultural boundaries.

Author Keywords

Emotion detection, affect detection, French, multi-lingual collaboration, distributed teams, cross-cultural communication, machine learning, natural language processing, qualitative analysis

ACM Classification Keywords

H.5.3. Group and Organization Interfaces: Computer-supported cooperative work, evaluation methodology.

General Terms

Human Factors; Experimentation; Measurement.

INTRODUCTION

Studying communication within distributed, multicultural teams has been one of the most established foci of interest within the CSCW community (eg, [1,6,7,8]), and is an increasingly important topic as more and more engineering, design, commerce, and other work groups are distributed around the globe. This avenue of inquiry has long posed challenges to research methodologies within CSCW. Neale et al have identified challenges involved in evaluation of CSCW systems [3], which only grow in severity with an increasing focus on analyzing traces of activity and communication, such as chat logs, directly and at scale.

In submission to CSCW2014

Studies in the last twenty years have increasingly evidenced a growing interest in understanding emotion and affect in the workplace, as factors influencing performance and dynamics in cooperative work environments [5]. This trend, combined with the increasing volume of text-based communication – a rich format for understanding affective processes within groups – has led to an upsurge in research on affect detection in text, including work in fields as diverse as sentiment analysis, affective computing, linguistics, and psychology, among others [5].

There are many challenges facing automating affect detection that are particular to chat [2]. Multilingual communication presents an additional set of problems, as not only are the words in another language, but the sentence structure, slang, idioms, and cultural context can be vastly different. An English-based affect classifier on French text would be heavily biased, and respond only to language-agnostic signals like punctuation, capital letters, and smileys. It is not sufficient to translate the messages and run an English classifier on them: automated translation fares poorly with informal text, and poorer still with the non-standard communication present in chat. Furthermore, quirks in French, such as the use of *bon* and *bien* (*good*) as a filler word [10], or the tendency to phrase things in the negative more frequently [4], mean literally translated French could throw off a classifier.

At the same time, multilingual chat logs present a unique and exciting opportunity for learning about group dynamics in multicultural collaborative work, which is of interest to the CSCW community.

In this manuscript, we focus on four years of chat logs from the Nearby Supernova Factory, an international team of astrophysicists located in America and France. They operated their telescope remotely three nights a week, with chat as their primary means of communication. The collaborative tools available to the scientists (see Fig. 1 for a screenshot of the control window chat room) were designed to help bridge cultural differences [1]. English was the main language spoken in the chat room, but when only native French speakers were working they tended to switch to French. The chats were used for strategizing and decision making, sometimes urgent trouble-shooting, and sometimes joking or chatting about non-work related things.

There are a total of 485,045 messages in the dataset, about a quarter of which are in French (although this number is hard to quantify, especially given the degree of language mixing, sometimes within a single line of chat).

We make use of these logs to build on prior work in supporting statistical affect detection in chat logs [2] by extending ALOE with feature extraction mechanisms particular to French. In this paper, we answer the following three questions:

> *Q1:* What specific feature extraction mechanisms improve accuracy on French?
>
> *Q2:* Is it valuable to distinguish an utterance as French rather than English, as a way to aid affect detection?
>
> *Q3:* Can these features make distinguishing French from English more accurate?

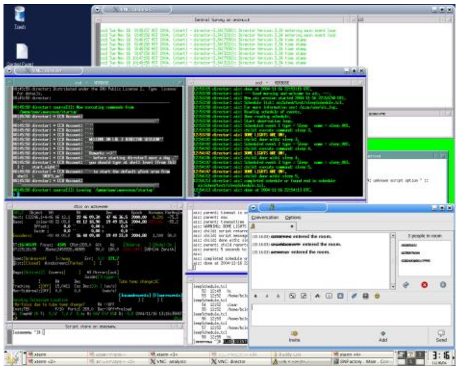
OUR APPROACH

ALOE trains on segments of chat manually labeled with emotion, and extracts seven types of features from its chat segments: message metadata (such as message length and frequency), pronouns, grouped word sets (known people's names, swear words, and negation words), character-level features (such as repeated letters or capital letters), emoticons, punctuation, and bag of words.

Most of these apply to French as well as English utterances: metadata is wholly language-agnostic, for instance. However, many of these features need to be adapted to be suitable for a different language, such as adding the translations of keywords, and allowing for the multiple transformations of spellings that result from text chat as a communication medium. For instance, "moi-même" is the correct way to say "myself", but it would more frequently be written "moi-meme" or "moi meme" in a chat setting. Similarly, in formal French "je" ("I") gets shortened to "j' " before a verb starting with a vowel, and in casual French "j'ai" ("I have") can get shortened to "jai" or the letter "g", which is pronounced the same way. All of these are allowed for in the French pipeline, which extends the existing ALOE codebase with several new feature extraction mechanisms.

Special strings presented a slightly larger challenge. No change was needed for the filter of known people's names, and for swear words we simply added a list of popular French swear words to the English filter.

Coding in negation words was more challenging, since negation is more complicated in French than in English. The word "not" is two words, "ne" and "pas", which come before and after the verb respectively. However, "ne" is abbreviated to "n' " before verbs beginning with a vowel, "pas" can be replaced with a number of other words (some of which have meanings besides the negative one), and in informal chat speak, either of these could be omitted, or the



Stef: je te laisse faire emile... Pour gagner du temps, je m'appretais a taper le clidup ;)
Translation: i'll let you do it emile... To save time, i'm getting ready to type the clidup ;)

Marcel: OK, we should have the telescope once the previous observer is finished ...

Stef: euh, you know, except bert everybody is french here... No need to speak english ! :-D

Pascal: Not true ...

what you typed is logged , and can be used on the long therm at the best source of log

Stef: ok

Pascal: at =as

main bon ... si on peut plus dire de c****** ... c'est nul
Translation: but then again ... if we can no longer say bull****, what's the point*

*sic: "main" means hand, "mais" means but.

Figure 1. *Top:* the desktop of a member of the *SNFactory* collaboration, with the chat room seen in the lower-right corner. *Bottom:* an exerpt of the chat log, complete with self-awareness, misspelling, switching languages and topics, smileys, and swearing.

pride serenity	amusement joy	ecstasy
agreement	supportive	gratitude
acceptance	trust	admiration
tired distraction	disbelief surprise	amazement
considering	relief	excitement
interest	anticipation	vigilance
apologetic pensiveness	embarrassment sadness	grief
apathy boredom	frustration	disgust
disagreement apprehension	confusion fear	terror
annoyance	impatience anger	rage

Figure 2. For each message in the chat log, human annotators determined which, if any, of the following codes applied. This taxonomy extends the Plutchik taxonomy of emotion for text chat [9]

apostrophe could be omitted so that “n” just becomes part of the verb. We allowed for as many of these cases as possible by including “ne”, “n’ “, “pas”, and most of the possible substitutes for “pas” – all the ones which tend to be used in a negative context, as well as many popular verb conjugations beginning with vowels, with “n” tacked on to the beginning.

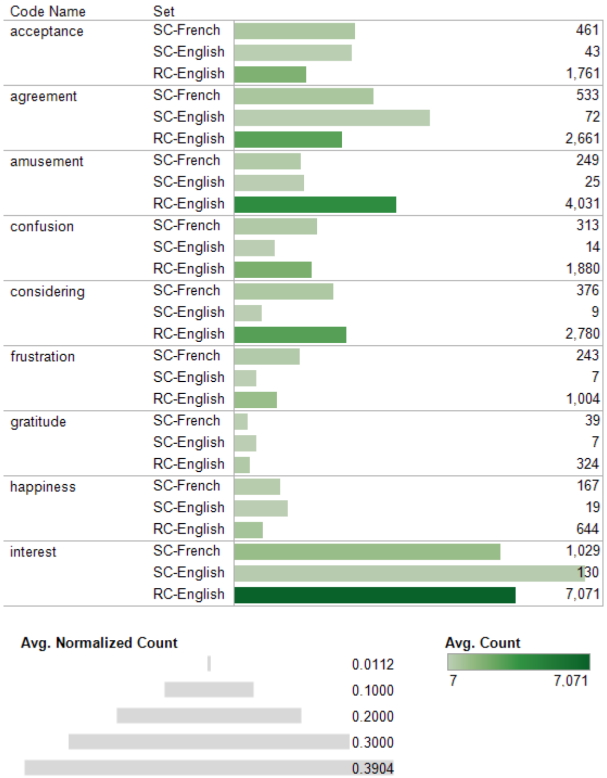


Figure 3. Code application distribution across three datasets: *SC-French*. *SC-English*, and *RC-English* (the first two were coded by a single French-speaking coder, the third – by the rest of the ~10 coders). The length of bars reflects a normalized count (proportion of that code application across all code applications *within that dataset*), whereas the coloration and right-hand label reflects the raw count of instances of messages labeled with that particular code.

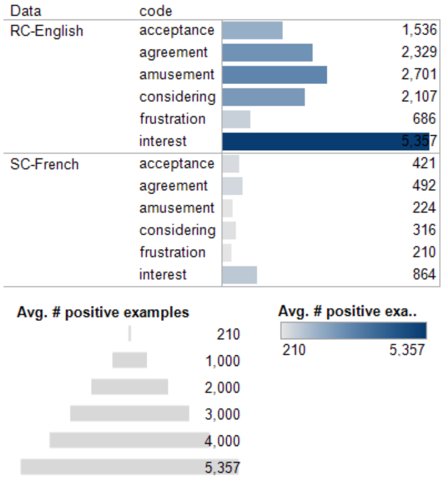


Figure 4. Number of positive examples of each of the top 5 codes in the SC-French set, and the number of positive examples for each corresponding code in the SC-English set.

Low-level spelling features include number and length of capital letter segments, number and length of “hm” variants, number and length of laughter phrases (“lol”, “haha”, “hehe”, etc), and length of repeated letter segments. Capitalized letters and repeated letters needed no change. We added “hum” and “heu”, French variants of “hm”, to the “hmm” filter. We also added “mdr”, the French equivalent of “lol”, to the laughter filter, even though in practice the francophone chatters tended to use “lol” much more.

Emoticons and punctuation needed no change, and we got rid of the English stemmer in the bag of words. We also added two features for words associated with acceptance (“d’accord”, “ok”, “okay”, etc.) and agreement (“agree”, “yes”, “oui”, etc.).

METHODOLOGY

We used three datasets to answer the three questions: *SC-French* (single-coder French), *SC-English* (single-coder English), and *RC-English* (remaining-coders English). The first author coded over 7k messages in the dataset for affect using a taxonomy (Fig. 2) developed for this dataset [9], most in French, and some in English, which comprise the *SC-French* and *SC-English* datasets, respectively. The set of all coded data by the remainder of the coders, which was considerably larger and on which original tests of ALOE were performed [2], comprised the *RC-English* set. Figure 2 shows the distribution of code application across the three datasets. The potential ramifications of these distributions with regard to the results are discussed in the Discussion section. Figure 4 shows also the number of positive examples for each of the top six emotion codes. ALOE works by building a binary classifier for each emotion, so those emotions with too few positive examples will be associated with poor performance. For the following experiments, we focused on the top six codes that each had more than 200 positive examples in the SC-English dataset.

Part of the challenge of working with a trace of a multilingual expert team is the relative difficulty of identifying language use. While most of the log is in English, many messages are in French; however, due to misspelling and (English) jargon common to this kind of dataset, reliable determination of language is not easy. Based on our results, we suggest that the enhanced ALOE pipeline can actually be used to reliably determine language in future work. For developing the sample of logs to code for the SC-* sets, however, we counted the number of messages including the word *le* as a coarse proxy for retrieving the top two dozen French logs.

We ran a series of experiments comparing the *Original* ALOE pipeline as reported in [2], and the modified *French* pipeline which included all the original feature extraction mechanisms as well as the additional mechanisms described in the Our Approach section. For each of the top six emotion codes, we varied the dataset (*SC-French* vs. *RC-English*), pipelines (*Original* vs. *French*), and ALOE's stemmer parameter (whether to use an English stemmed, a French stemmer, or no stemmer).

In each experiment, we used one of the datasets, one of the pipelines, and some configuration of ALOE parameters to train and test a classification model, producing an F-measure we could use to compare performance across experiments. For training, *RC-English* data was down-sampled. The number of positive examples (Fig. 4) is fewer than the raw number of coded messages (Fig. 3) due to ALOE's segmentation mechanisms.

Each model was tested with a balanced set of the same number of positive and negative examples, in a 10-fold cross-validation procedure. During this, a tenth of the data was held out from the training set, and a different such small subset was left out in each of ten runs that produced a result that could be evaluated using the F-measure. This measure combines precision and recall; an F-measure of 1 characterizes a perfect classifier than makes no misclassifications.

Each F-measure shown in Figure 5 is the average of 10 runs during cross-validation, a typical approach to evaluating a classifier training procedure intended to be robust to outliers in the data. In addition to using F-measure to evaluate models, we manually examined misclassified data-points to gain insight into possible sources of error and avenues for improvement.

All code used during experimentation has been added to the publicly-available ALOE codebase <and will be made visible upon acceptance of this manuscript>. The resulting measurements in greater detail, including experimental runs that are not included in this report, are available for download and examination at <URL redacted for anonymity>. Below, we highlight the results we believe to be relevant to the CSCW community.

RESULTS

In general, the French pipeline performed as well as or better than the original on the six more frequent affect codes across both English and French datasets. The results for *Data* x *Pipeline* experiments for each code are shown in Figure 5, with the corresponding top-weighted features on the *SC-French* dataset shown in Figure 6. The feature weights produced by an SVM do not reflect causal relationships between observed features and true classifications, but they do provide a window into the workings of the model that may inform choices about feature selection, crucial in the context of this kind of subjective classification task [2].

The two pipelines performed similarly on *interest*; the corresponding top-weighted features are, in both cases, question marks and length of question marks, which are language-agnostic features. The same goes for *amusement*, where the top features are generally emoticons and laughter, also language–agnostic. However, the French pipeline performed significantly better on *acceptance* and *agreement*, making use of the new features targeted at those codes. This improvement was, unsurprisingly, accentuated in *SC-French*, and did not reduce performance on *RC-English*. Upon manual inspection, we found that many false negatives for amusement included smileys (":)"), which the single coder had tended to code as happiness, except in response to a joke. Likewise, a large portion of the false negatives for happiness were winking smileys, which tended to be coded as amusement.

There was no notable difference due to pipeline on

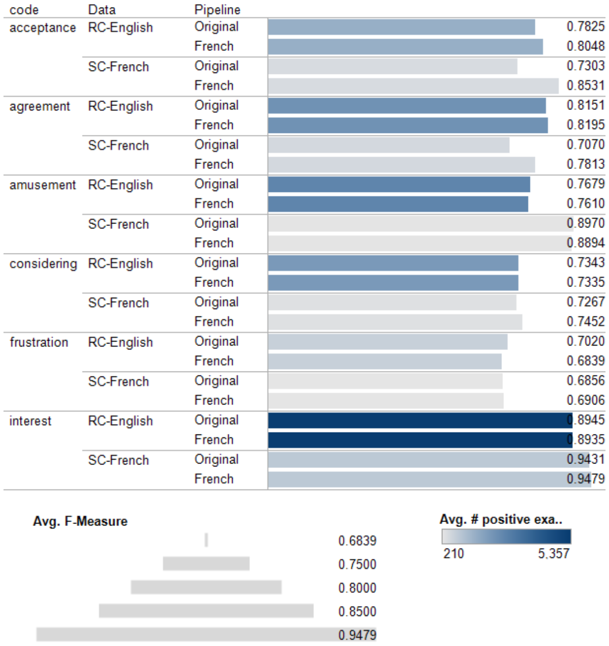
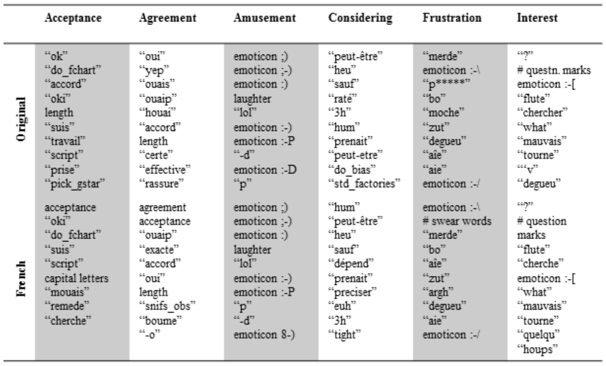


Figure 5. F-measures, which take into account both precision and recall, of classifiers trained using each of the two different pipelines on balanced test sets for each of the codes. The number of positive examples is encoded in the color of each bar.



	Acceptance	Agreement	Amusement	Considering	Frustration	Interest
Original	"ok" "do_fchart" "accord" "oki" length "suis" "travail" "script" "prise" "pick_gstar"	"oui" "yep" "ouais" "ouaip" "houai" "accord" length "certe" "effective" "rassure"	emoticon ;) emoticon ;-) emoticon :) laughter "lol" emoticon :-) emoticon :-P "-d" emoticon :-D "p"	"peut-être" "heu" "sauf" "raté" "3h" "hum" "prenait" "peut-etre" "do_bias" "std_factories"	"merde" emoticon :-\ "p*****" "bo" "moche" "zut" "degueu" "aîe" "aie" emoticon :-/	"?" # questn. marks emoticon :-["flute" "chercher" "what" "mauvais" "tourne" ""v" "degueu"
French	acceptance "oki" "do_fchart" "suis" "script" capital letters "mouais" "remede" "cherche"	agreement acceptance "ouaip" "exacte" "accord" "oui" length "snifs_obs" "boume" "-o"	emoticon ;) emoticon ;-) emoticon :) laughter "lol" emoticon :-) emoticon :-P "p" "-d" emoticon 8-)	"hum" "peut-être" "heu" "sauf" "dépend" "prenait" "preciser" "euh" "3h" "tight"	emoticon :-\ # swear words "merde" "bo" "aîe" "zut" "argh" "degueu" "aie" emoticon :-/	"?" # question marks "flute" "cherche" emoticon :-["what" "mauvais" "tourne" "quelqu" "houps"

Figure 6. Top features for each of the classification models trained across six emotions and two pipelines on the *SC-French* dataset.

frustration or *considering*. The slight decline in accuracy on frustration in *RC-English* could be the result of a number of changes – for instance, we took out the stemmer in the bag of words feature to make the new pipeline work equally well on French and English. However, that lack of specificity may give the original pipeline, with the English stemmer, an edge in *RC-English*, which could be exaggerated on frustration, a label that has many words that are strongly positively correlated with it.

In some cases, the additional features improved performance on *SC-French*; in no case did the French pipeline significantly reduce classification quality on the English dataset. We also experimented with more language-specific parameters by varying the stemmer used by ALOE, finding that on the French dataset, the absence or presence of the French stemmer made no difference, consistent with the findings in [2] that English stemmer does not improve the quality of affect detection.

This highlights the language-agnosticism of the existing ALOE pipeline, and does not necessarily rule out the potential benefit of identifying which language is spoken in a chat log – itself a potentially valuable step in data preparation. We can use the same machine learning approach to tackle the challenge of language detection in a very noisy set. Surprisingly, the French pipeline performed no better than the original, though both performed well, with an F-measure of 0.93. This was an exciting finding, as French and English frequently appear in the same message alongside jargon, which prevented us from effectively using other non-ML methods of language identification.

DISCUSSION

Some of the error in classification is due to cross-contamination. Happiness (joy) in the taxonomy of emotion is theoretically very near amusement (Fig. 2), so the single coder's bias that may account for error in the *amusement* code, as we speculate in the Results section, is neither surprising nor easily avoided before the damage is done. Below, we expand in the idea of legitimizing coder bias as an inherent aspect of a subjective classification process, and some possible implications this has for the use of automation tools like ALOE enables.

A limitation of this work is that only the first and last authors speak French. The single French coder differed in the distribution of affect code applications from the remainder of the coders. Figure 4 shows that the distribution in *SC-French* and *SC-English* were generally consistent with the distribution in *RC-English*, with some exceptions. *SC-English* and *SC-French* had a much smaller proportion of *amusement* then *RC-English*, and a slightly larger proportion of *acceptance* and *agreement*. Some variation is unsurprising because identifying affect is a subjective task; this variation alone is not unusual in the rest of the dataset and does not undermine the validity of the coding as measured by inter-rater reliability [2]. These variations may be based on a different interpretation of the meaning of labels that are not only inherently subjective, but also culturally-situated. For instance, the single coder tended to have a narrower notion of amusement, limited mostly to laughter and smileys, while the remaining coders also positively identified *amusement* when people were joking around ("Too much work for a day off already!" or "If anyone does that, they should be thrown off the mountain, from a trebuchet."). This difference on the one hand means there are fewer positive examples in *SC-French* for the classifier to train on, but could also mean that amusement is more consistent in *SC-French*, which would lead the classifier to perform better. The disparity in *acceptance* and *agreement*, on the other hand, is potentially due to the fact that when a message could possibly be read as expressing either acceptance or agreement (eg, 'ok'), the single coder tended to mark it as both, whereas the remaining coders may have tended toward picking one. This could possibly translate to more consistency for these codes in *SC-French*, which could help the classifiers compensate for the smaller sample size.

Some of the results suggest that the consistency of a single labeler may help an approach like ALOE effectively mimic that coder's behavior. In the original ALOE publication, Brooks et al argued that the automation of emotion detection can be validated by having multiple coders code the same chunks of data, calculate reliability of the overlap, and if the reliability is acceptable, train classifiers using the data generated by all coders – even that which is only coded by a single coder [2]. We suggest an alternative route, training a classifier on a single coder, may be more effective. While this approach is more subject to the biases of a single coder, it is more robust to differences in coder biases that may result in weaker classifiers confused by increased noise in the definition of a particular code.

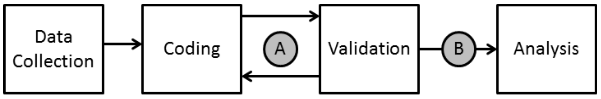


Figure 7. In future work, we suggest exploring changing the placement of the automation intervention to be earlier in the qualitative data analysis pipeline (A) rather than later (B), contrary to the assumptions in [2].

Cultural context, not only bias and noise due to subjectivity, poses a challenge. The single coder is not from France, which means there may be cultural nuance that she missed, or read in a different way than was intended. It's possible that a French person reading the same logs would identify different emotions and affect. While this is a limitation, it's not avoidable (and not necessarily a large problem) for two reasons. Even within a country, cultures are different from community to community, and the same utterance could have different nuance depending on the person. Therefore, while a French coder may have more context than a French-speaking American, there will never be a coder who has the entire context except for the author of the message.

Secondly, in the context of a geographically distributed group, a person's intended meaning or emotional value doesn't necessarily matter more than the meaning perceived by the rest of the group.

CONCLUSION

We contribute an update to an English-specific classifier that makes it applicable to both French and English; all code and results are available publicly online for use and comparison[1]. We have been able to answer three questions involving multilingual affect detection as follows:

Q1: What specific feature extraction mechanisms improve accuracy on French? The French pipeline does not hamper the performance of ALOE on English, while providing some benefit on French utterances, particularly on *agreement* and *acceptance* codes, potentially due to benefits from the agreement-related and acceptance-related sets of combined English and French cues.

Q2: Is it valuable to distinguish an utterance as French rather than English, as a way to aid affect detection? No: the only language-specific mechanism of ALOE, the optional stemmer, did not provide any benefit.

Q3: Can these features make distinguishing French from English more accurate? No, although both the original and the French pipelines perform very well on this task.

Understanding communication and social interaction within distributed multi-cultural teams has long been of central relevance to the CSCW community. In his 1990 Whole Earth Review article, Hiroshi Ishii writes (emphasis added):

> …The world is full of communities that differ in fundamental ways. However, we live together on the planet, and we must collaborate internationally. We work together in scientific research, cooperate and compete economically, and discuss problems of mutual interest. Yet we know very little about the dynamics of cross-cultural communication, and little about the unique cultural biases in the way we communicate and make decisions. **… We should use all the tools at our disposal to pay more attention to understanding the differences among us.** And then we should start to think how to overcome this gap with or without technology.

Our work seeks to make it possible to analyze large multilingual chat logs to better understand team dynamics externalized in those traces. Additionally, we suggest, in the Discussion section, possible ramifications of studying subjective phenomena, such as affect, in a multicultural context where even manual coding is inherently noisy. Specifically, the practice of validating coding prior to coalescing all coded data for training classification (as in [2]) may be less appropriate than integrating automation into an *individual coder's* workflow. In this scheme, illustrated in Fig. 7, each individual coder's work is used as training for classifiers, which are then validated using reliability metrics, and the combined human and automatic labeling are then used for further analyses. We intend to explore this process in future work, as well as other possible strategies to enable human annotators from different cultural backgrounds, with potentially different understanding of coding schemes, to help contribute to interpreting data without undermining the applicability of automation and its benefit in a data-rich setting.

ACKNOWLEDGMENTS

Redacted for anonymity

REFERENCES

1. Cecilia R. Aragon, Sarah Poon. No Sense of Distance: Improving Cross-Cultural Communication with Context-Linked Software Tools. iConference 2011.
2. Brooks, M., Kuksenok, K., Torkildson, M. K., Perry, D., Robinson, J. J., Scott, T. J., Anicello, O., Zukowski, A., Harris, P., Aragon, C. R. 2013. Statistical Affect Detection in Collaborative Chat. CSCW 2013.
3. Dennis C. Neale, John M. Carroll, Mary Beth Rosson. Evaluating Computer-Supported Cooperative Work: Models and Frameworks. CSCW 2004.
4. French Discourse: Toward an Accrued Negativization
5. Grandey, A. Emotions at Work: A Review and Research Agenda. In Handbook of Organizational Behavior. SAGE, London, 2008.
6. Hao-Chuan Wang, Susan R. Fussell, Leslie D. Setlock. Cultural Difference and Adaptation of Communication Styles in Computer-Mediated Group Brainstorming. CHI 2009.
7. Hiroshi Ishii. Cross-Cultural Communication & Computer-Supported Cooperative Work. Whole Earth Review, 1990.
8. Leslie D. Setlock, Susan R. Fussell, Christine Neuwirth. Taking It Out of Context: Collaborating within and cross Cultures in Face-to-Face Settings and via Instant Messaging. CSCW 2004.
9. Taylor Scott, Katie Kuksenok, Daniel Perry, Michael Brooks, Ona Anicello, Cecilia Aragon. Adapting Grounded Theory to Construct a Taxonomy of Affect in Collaborative Online Chat. SIGDOC 2012.
10. The French Experiment. "French Conversation Fillers." http://www.thefrenchexperiment.com/learn-french/fillers.php. Retrieved May 31, 2013.

[1] URL redacted for anonymity