Towards Automatic Annotation of Clinical Interview Transcripts

Alexander Kotov
Department of Computer Science
Wayne State University
kotov@wayne.edu

April Carcone
Pediatric Prevention Research Center
Wayne State University
acarcone@med.wayne.edu

Ming Dong
Department of Computer Science
Wayne State University
mdong@wayne.edu

Sylvie Naar-King
Pediatric Prevention Research Center
Wayne State University
snaarkin@med.wayne.edu

Kathryn Brogan
Department of Dietetics and Nutrition
Florida International University
kabrogan@fiu.edu

ABSTRACT

This work addresses the problem of automatic annotation of clinical interview transcripts with semantic categories. We formulate this task as supervised machine learning problem and propose scalable and efficient probabilistic classifiers based on generative latent variable models to solve it. Experimental results indicate that the proposed classifiers outperform popular standard classification models, such as Naïve Bayes, while providing interpretable results for clinicians or healthcare researchers.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Health; I.2.7 [**Natural Language Processing**]: Text analysis

Keywords

Electronic Health Records, Machine Learning, Textual Classification, Graphical Models

1. INTRODUCTION

The problem of assigning codes from a predefined code book to textual data fragments (such as entire documents or their parts) arises in many contexts. Such codes can be viewed as semantic labels or abstractions from raw textual data. Besides simple cataloging, such abstractions can facilitate many tasks in textual data analysis, including analysis of electronic health records in the form of patient interview transcripts. Dialog between a caregiver and a patient is an integral part of clinical studies and interventions for many conditions and disorders. In this work, we focus on childhood obesity and use the transcripts of motivational interviews conducted at a Pediatric Prevention Center (PPC) as our dataset. In particular, as part of the treatment process, PPC clinicians conduct interviews with children and their parents to discuss their lifestyle, diet and eating habits in order to identify the root cause of obesity. Detailed analysis of those interviews aims at identifying communication strategies that are effective in triggering patients' commitment to behavior change that will ultimately lead to weight loss. Such strategies are typically derived through retrospective analysis of past interview transcripts. Part of this analysis is assignment of codes to patient replies during the interviews. Analyzing sequences of assigned codes allows clinicians to better understand the patient's thought process during the course of the interviews, without having to wade through entire transcripts over and over again. Such understanding in turn leads to identification of the mechanisms of effect for intervention models, which can then be used to refine theory and guide clinical practice.

Transcript coding has traditionally been performed manually by trained coders, which is a tedious and resource intensive process. Therefore, methods that can efficiently and accurately distinguish the nuances of patient-provider interactions in interview transcripts can have a tremendous positive impact on many areas of clinical practice and research. Inferring psychological state of the patients during interviews using only lexical content of their transcriptions is a challenging task for several reasons. First, many important indicators of emotions such as gestures, facial expressions and intonations are lost during the transcription process. Second, some utterances from patients during the interview may be too short and lack sufficient context for accurate classification. Furthermore, patients come from a variety of educational backgrounds and their use of language may be different. This problem is exacerbated when the interviews are conducted with children and adolescents, since children in general tend to often use incomplete sentences and frequently change subjects. This work constitutes the first step towards solving the problem of automatic coding of electronic health records (EHR) in the form of interview transcripts.

In particular, in Section 2 we provide a brief overview of the previous work related to this study, in Section 3 we present the details of the proposed probabilistic classifiers

KDD-BHI'14, August 24, 2014, New York, New York.

and compare their accuracy with standard classifiers, such as Naïve Bayes and Support Vector Machine, in Section 4.

2. RELEVANT WORK

The task of fine-grained analysis of language arises in many contexts from sentiment analysis of on-line reviews [7] to analysis of scientific literature [1] and digital forensics [2]. Performance of different classifiers (including Naïve Bayes [5]) on most common text classification tasks, such as sentiment analysis [4] [6], has been examined in detail in previous work. However, most of these classification tasks are binary. A method to automatically assign semantic types to dialog turns in transcripts of phone conversations between nurses and dialysis patients was proposed in [3]. While this work explored the effectiveness of different strategies of leveraging external resources either by replacing some of the terms in bag-of-words feature vectors with concepts from the Unified Medical Language System (UMLS) or by representing the transcript fragments in the space of clusters derived from a large external corpus, it did not compare the performance of different classifiers. Also, the transcripts used in that work contained large amount of medical terminology, which made using specialized external resources possible. This work presents the results of initial experiments with different classification models for automatic annotation of non-technical clinical transcripts, such as those of motivational interviews, with fine-grained semantic types and without leveraging external resources.

3. METHODS

3.1 Classes

In this work we classify the fragments of interview transcripts into the following 5 language categories, which are summarized in Table 1:

CL-: negative commitment language
CL+: positive commitment language
CT-: negative change talk
CT+: positive change talk
AMB: ambivalence

The language categories (classes) in Table 1 represent the target patient behaviors, on which clinicians are focused when conducting motivational interviewing, the specific intervention strategy used in the clinical encounters analyzed in this study. When coding using traditional qualitative coding techniques to identify these target behaviors, two coders are typically employed. The primary coder codes all sessions, while the secondary coder codes 20% of randomly selected sessions to assess inter-rater reliability. Coders used the "Minority Youth-Sequential Code for Observing Process Exchanges" (MY-SCOPE) coding manual to assign the code to each patient utterance in the interview transcripts used for this work. MY-SCOPE defines the five major classes of patient utterances presented in Table 1, among others. Change talk (CT) is statements describing patients' own desires, abilities, reasons, and need for adhering to diet recommendations and are positive, when supportive of behavior change, and negative, when against behavior change. Commitment language (CL) is statements about their intentions or plans for adhering to desires and intentions, which again, are positive, when supportive of behavior change, and negative, when against behavior change. Ambivalent utterances (AMB) are change talk or commitment language statements that contain a combination of positive and negative sentiments about changing one's behavior. In the following sections, we present the classifiers used for the task of differentiating the above classes as well as their combinations and report their accuracy.

3.2 Baselines

We formulate the task of automatic coding of interview transcripts as a supervised machine learning problem. We use standard bag-of-words feature generation framework, in which a predefined set of lexical features (vocabulary), $V = \{w_1, \ldots, w_N\}$, can appear in a given textual fragment. For example, one such feature could be the number of times a word (unigram) "exercise" appears in a given textual fragment. This way each textual fragment f is represented as a feature vector $\langle n_{w_1,f}, \ldots, n_{w_N,f} \rangle$, where $n_{w,f}$ is the number of times feature w occurs in f. To determine the best classification model for this task, in the following sections we experimentally compare standard supervised machine learning algorithms, such as Naïve Bayes (NB) classifier and Support Vector Machine (SVM), with our proposed probabilistic classification models.

3.2.1 Naïve Bayes

Naïve Bayes is a standard probabilistic classifier, which given a textual fragment $f = \{w_1, \ldots, w_{N_f}\}$ consisting of N_f words assigns it to the class c^*, such that $c^* = argmax_c p(c|f)$, where $p(c|f)$ is estimated by applying the Bayes' rule as follows:

$$p(c|f) = \frac{p(f|c)p(c)}{p(f)} \propto p(f|c)p(c)$$

In order to estimate $p(f|c)$, Naïve Bayes classifier makes an assumption about conditional independence of features given c, the class of fragment f:

$$p(f|c) = \prod_{i=1}^{N_f} p(w_i|c)^{n_{w_i,f}}$$

Despite its relative simplicity, NB has been experimentally demonstrated to be one of the most effective text classification algorithms. In this work, we used a standard implementation of Multinomial NB algorithm from the Weka text mining toolkit [1].

3.2.2 SVM

Support Vector Machine is a geometric (large-margin) classifier that has been shown to be highly effective for classification tasks on a variety of data types (not only textual). In case of two-class (binary) classification, given a set of training data samples, SVM finds a hyperplane, represented by a weight vector $\vec{w}$ and intercept b, that separates the feature vectors in two classes in such a way that the margin (the distance from the closest samples in each class to the hyperplane) is maximized. Classification of test samples consists of determining which side of the hyperplane they fall on:

$$c = sign(\vec{w}^T \vec{f}) + b$$

SVM is often used along with a kernel function that transforms the feature space. In this work, we used a standard implementation of SVM algorithm with polynomial kernel from the Weka toolkit.

[1] `http://www.cs.waikato.ac.nz/ml/weka/`

Category	Example
CL-	"With the food choices like sometimes when we're on the go, we go pick something up, not at home. Like when I'm in town with my dad, he doesn't cook that much so we eat out a lot when I'm with him"
CL+	"I just want to not eat so much of it in one day. Maybe have, like, one meal from a fast food restaurant or probably like, three times a week or something like that"
CT-	"I'm not very educated with the calories in the different foods that I eat. I really don't know like especially going out fast food eating what the calories are in (some of the food)"
CT+	"I like all these ideas, I really do , because I like doing the, at the YMCA where we went there was a like a crunch machine where you do crunches. You sit down on it and then you pull up and stuff. I like doing that and it gave me more strength and I like weights too"
AMB	"Yeah, I think maybe I'm going to try it out. I don't know how long it's going to last though. Like once I actually get started, okay, I'm done with this. Might sit in the back of my closet, likes the karate uniform.

Table 1: Language categories considered in this work

3.3 Probabilistic models

We propose two probabilistic generative latent variable models for the task of automatic coding of interview transcripts: Latent Class Allocation (or LCA) and Multi-Granularity Labeled Latent Dirichlet Allocation (or MG-LDA). LCA associates a latent variable with each word that determines its type (whether a word is general or characteristic of a certain class). MG-LDA is an extension of LCA, which besides a latent variable determining the type of a word associates with each word another latent variable determining its class-specific category or topic. Thus, MG-LDA is a *structured* version of LCA.

3.3.1 LCA

LCA models each textual fragment f labeled with class c_f as a set of draws from a mixture distribution λ_f of a background multinomial ϕ^{bg} drawn from a Dirichlet prior β^{bg} and a multinomial ϕ^{cls} specific to c_f. Background and class-specific multinomial distributions are over the entire vocabulary V (also referred as language models (LMs)). LCA generates the textual fragments in clinical interviews according to the following probabilistic process:

1. draw $\lambda_f \sim Beta(\gamma)$, a binomial distribution controlling the mixture of background and class-specific LMs for f
2. for each word position i of N_f in f:
   (a) draw Bernoulli switching variable $m_{f,i} \sim \lambda_f$
   (b) if $m_{f,i} = bg$:
      i. draw a word $w_{f,i} \sim \phi^{bg}$
   (c) if $m_{f,i} = cls$:
      i. draw a word $w_{f,i} \sim \phi^{cls,c_f}$

The generative process of LCA in plate notation is illustrated in Figure 1.

Classification using LCA is done using class-specific multinomials ϕ^{cls} (or $p(w|c)$) to derive $p(c|w)$:

$$p(c|w) = \frac{p(w|c)p(c)}{p(w)}$$

where $p(c) = \frac{n_{f,c}}{M}$ ($n_{f,c}$ is the number of interview fragments labeled with class c and M is the total number of fragments) and $p(w)$ is a probability of word in a collection language model estimated using maximum likelihood. $p(c|w)$ is used to directly classify f:

$$c^* = argmax_c p(c|f) = \prod_{i=1}^{N_f} p(c|w_i)^{n_{w_i,f}}$$

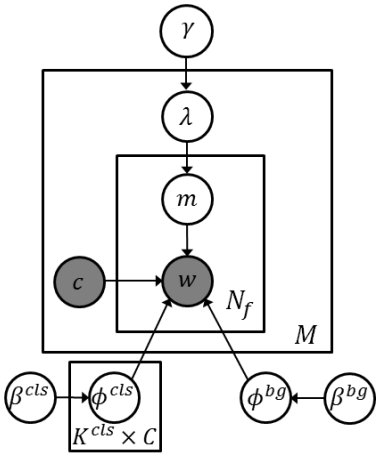

Figure 1: Graphical model of LCA

3.3.2 MG-LDA

MG-LDA models each textual fragment f labeled with class c_f as a mixture of the background topic (multinomial) ϕ^{bg} drawn from a Dirichlet prior β^{bg} and K^{cls} topics (multinomials) drawn from a Dirichlet prior β^{cls}. MG-LDA generates the textual fragments in clinical interviews according to the following probabilistic process:

1. draw $\lambda_f \sim Beta(\gamma)$, a binomial distribution controlling the mixture of background and class-specific topics for f
2. draw $\Theta_f^{cls} \sim Dir(\alpha^{cls})$, a distribution of class-specific topics for f
3. for each word position i of N_f in f:
   (a) draw Bernoulli switching variable $m_{f,i} \sim \lambda_f$
   (b) if $m_{f,i} = bg$:
      i. draw a word $w_{f,i} \sim \phi^{bg}$

(c) if $m_{f,i} = cls$:

i. draw a topic $z_{f,i} \sim \Theta_f^{cls}$

ii. draw a word $w_{f,i} \sim \phi_{z_{f,i}}^{cls,c_f}$

The generative process of MG-LDA in plate notation is illustrated in Figure 2.

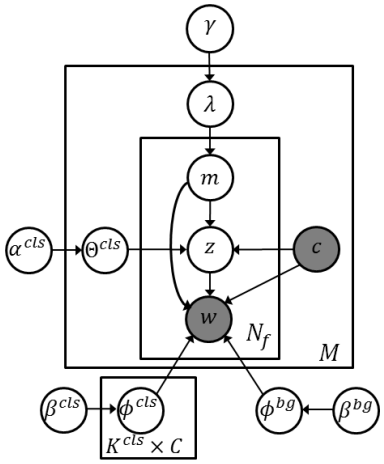

Figure 2: Graphical model of MG-LDA

Classification using MG-LDA is done by first deriving class-specific LM $p(w|c)$ per each class c from class-specific topics ϕ^{cls,c_f} (or $p(w|c,z)$) by marginalizing over z:

$$p(w|c) = \sum_{z=1}^{K^{cls}} p(w|c,z)$$

and using class-specific LMs to directly classify f as in LCA.

4. EXPERIMENTS

The experimental dataset used in this work consists of 1505 manually annotated fragments of interview transcripts. As follows from Table 2, experimental dataset has a significant degree of class imbalance.

Class	Samples	%
CL-	48	3.19
CL+	410	27.24
CT-	177	11.76
CT+	780	51.83
AMB	90	5.98
Total	**1505**	**100.0**

Table 2: Number of samples per class in experimental dataset

The dataset was first pre-processed by removing very frequently occurring terms (that occur in more than 25% of textual fragments). We also used the following pre-processing methods to study the effect of each pre-processing method on classification performance:

RAW: no preprocessing, original dataset is used;

STEM: Porter stemmer is applied to each term in the dataset to eliminate morphological variation;

STOP: stop-words are removed, but stemming is not applied;

STOP-STEM: both Porter stemmer is applied to each term in the dataset and stop-words are removed.

For all experiments we used randomized 5-fold cross-validation. The Gibbs sampler for inference in LCA and MG-LDA was run for 1000 iterations. Classification performance of Naïve Bayes, SVM and LCA on the task of differentiating 5 language categories when experimental dataset is pre-processed with different methods is summarized in Tables 3, 4 and 5 respectively.

Method	Recall	Precision	F1-Measure
RAW	**0.489**	**0.514**	**0.496**
STEM	0.473	0.5	0.48
STOP	0.455	0.485	0.465
STOP-STEM	0.448	0.474	0.458

Table 3: Performance of Naïve Bayes using different pre-processing methods

Method	Recall	Precision	F1-Measure
RAW	**0.565**	**0.548**	**0.554**
STEM	0.531	0.515	0.521
STOP	0.512	0.476	0.483
STOP-STEM	0.498	0.469	0.477

Table 4: Performance of Support Vector Machine using different pre-processing methods

Method	Recall	Precision	F1-Measure
RAW	0.512	0.495	0.496
STEM	**0.531**	**0.501**	**0.507**
STOP	0.468	0.464	0.464
STOP-STEM	0.441	0.442	0.439

Table 5: Performance of Latent Class Allocation using different pre-processing methods

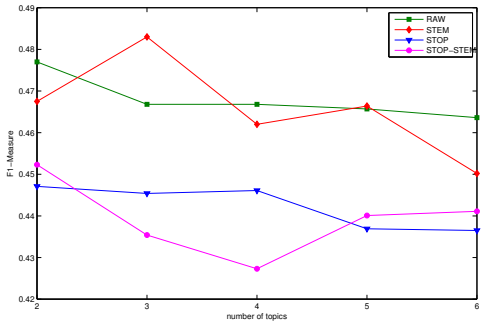

Figure 3: F1-Measure of MG-LDA by varying the number of topics and in combination with different pre-processing methods on the task of classification of all language categories

Since classification accuracy of MG-LDA is dependent on its parameter (the number of topics) we first optimized it with respect to F1-measure. Figure 3 indicates that the optimal classification results for MG-LDA in combination with different pre-processing methods are achieved when the

number of topics is small (2 or 3 in most cases). Classification performance of MG-LDA on the task of differentiating 5 language categories when experimental dataset is pre-processed with different methods is summarized in Table 6.

Method	Recall	Precision	F1-Measure
RAW	0.503	0.473	0.477
STEM	**0.521**	**0.474**	**0.483**
STOP	0.453	0.443	0.447
STOP-STEM	0.463	0.445	0.452

Table 6: Performance of Multi-Granularity Latent Dirichlet Allocation using different pre-processing methods

Several interesting conclusions can be derived from Tables 3, 4, 5 and 6. First, stemming and stop-word removal degrades classification performance or Naïve Bayes and SVM, while LCA and MG-LDA achieve the best classification performance when stemming is used. However, for all 4 classifiers used in this work stop word removal by itself and in combination with stemming decreases the accuracy of classification, which suggests that common stop words might be important indicators for some of the language categories. Second, as follows from Table 8, LCA outperforms Naïve Bayes in terms of F1-measure and recall, while MG-LDA only outperforms NB in terms of recall. This indicates that additional structure of class-specific multinomials introduced by MG-LDA does not translate into better classification accuracy. Third, although across all 3 classifiers SVM achieves the best performance of 0.554 with respect to the F1-measure, this model is much harder to interpret than LCA, which returns class-specific LMs directly. Examples of the most characteristic terms for each language category as determined by LCA are provided in Table 7.

class	words
CL-	eat food sometimes lot pop thirsty badly much drink sugar junk watch tv computer instinct fries sandwich pizza chips skip problem
CL+	walk play jog outside more salad run fast juice vegetables chicken water try go fruit lost snack little
CT-	laugh really cola splenda confused liter library joke hard not don't didn't ain't never tired nervous worried appetite
CT+	you can could more yeah want weight lose need healthy stuff better diabetes fun help cook probably happy track
AMB	know cousin grandmother grandma guess abandon cocktail kool aid milk people talk today anybody tofu

Table 7: Most characteristic words for each class according to LCA

As follows from Table 7, negative commitment language is strongly associated with the words reflecting bad diet ("drink", "sugar", "pop", "fries") and sedentary lifestyle ("watch", "tv", "computer"), while positive commitment language is strongly associated with the terms related to exercise ("walk", "play", "run") and healthy food options ("salad", "juice", "vegetables", "fruit"). The words characteristic or CT- and CT+ generally reflect the negative ("don't", "never", "tired") and positive ("can", "need", "lose", "happy") attitudes towards weight loss respectively.

Algorithm	Recall	Precision	F1-Measure
NB	0.489	0.514	0.496
SVM	**0.565**	**0.548**	**0.554**
LCA	0.531	0.501	0.507
MG-LDA	0.521	0.474	0.483

Table 8: Summary of the best performance of different classifiers on the task of distinguishing 5 original language categories. Best result for each metric is highlighted.

Algorithm	1	2	3	4	5
NB	0.336	0.436	0.213	0.631	0.247
SVM	0.308	0.504	0.261	0.698	0.252
LCA	0.066	0.452	0.228	0.675	0.122
MG-LDA	0.062	0.417	0.166	0.664	0.098

Table 9: Summary of the best per class performance in terms of F1 measure of different classifiers on the task of distinguishing 5 original language categories.

In the second set of experiments, we aggregated the interview fragments labeled as positive and negative commitment language (CL+ and CL-) and change talk (CT+ and CT-) into one combined class for commitment language (CL) and one combined class for change talk (CT) respectively and evaluated the accuracy of our classifiers in distinguishing the interview fragments labeled with the resulting three broader classes. As follows from Table 10, the distribution of broader classes exhibits imbalance and is skewed towards CT.

class	# samples	%
CL	458	30.43
CT	957	63.59
AMB	90	5.98

Table 10: Number of samples per aggregated positive and negative sub-classes of CL and CT

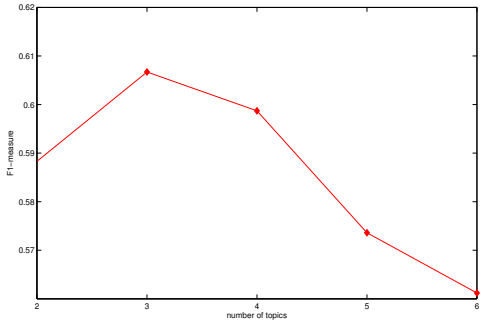

Figure 4: F1-Measure of MG-LDA by varying the number of topics on the task of classification of aggregated positive and negative sub-classes within CL and CT

We used the optimal pre-processing methods for each classifier (no pre-processing for SVM and NB and only stemming for LCA and MG-LDA). We also tuned MG-LDA with respect to F1-Measure and found out that again the optimal number of topics is 3. Performance of different classifiers on the task of differentiating the interview fragments labeled with CL, CT and AMB is summarized in Table 11. Again, LCA outperforms NB with respect to F1-measure, while SVM is slightly better than LCA.

Algorithm	Recall	Precision	F1-Measure
NB	0.595	**0.645**	0.613
SVM	**0.655**	0.644	**0.648**
LCA	0.652	0.634	0.641
MG-LDA	0.624	0.598	0.607

Table 11: Classification performance on aggregated positive and negative sub-classes within CL and CT. Best result for each metric is highlighted.

In the third set of experiments, we aggregated the interview fragments labeled as positive sub-classes of commitment language (CL+) and change talk (CT+) into one combined positive class (+) and negative sub-classes of commitment language (CL-) and change talk (CT-) into one combined negative class (-) and evaluated the accuracy of our classifiers in distinguishing the interview fragments labeled with the resulting modality-based broader classes. Similar to aggregating sub-classes *within categories*, aggregating sub-classes *across categories* also results in class imbalance (Table 12).

class	# samples	%
-	225	14.95
+	1190	79.07
AMB	90	5.98

Table 12: Number of samples per aggregated positive and negative sub-classes across CL and CT

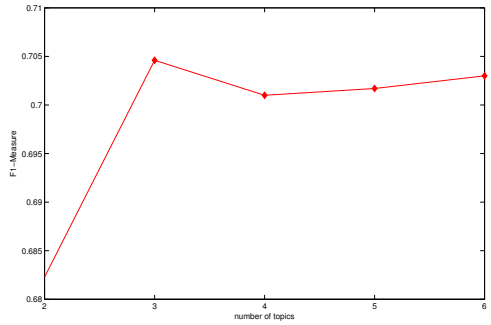

Figure 5: F1-Measure of MG-LDA by varying the number of topics on the task of classification of aggregated positive and negative sub-classes across CL and CT

As follows from Figure 5, MG-LDA achieves the best performance with respect to F1-Measure when the number of topics is 3. Performance of different classifiers on the task of differentiating the interview fragments labeled with +, - and AMB is summarized in Table 13.

Algorithm	Recall	Precision	F1-Measure
NB	0.649	0.718	0.675
SVM	**0.753**	**0.725**	**0.737**
LCA	0.714	0.7	0.706
MG-LDA	0.716	0.695	0.705

Table 13: Classification performance on aggregated positive and negative sub-classes across CL and CT. Best result for each metric is highlighted.

For this task, both LCA and MG-LDA outperform NB, while SVM results in the best overall F1-measure of 0.737, the highest for all 3 classification tasks.

5. SUMMARY AND FUTURE WORK

In this work, we proposed two novel, efficient and interpretable probabilistic classifiers, one of which outperformed standard Naïve Bayes on all classification tasks that we experimented with. This method was, however, unable to outperform SVM. Although automatic coding of clinical encounter data is a challenging problem, we were able to obtain promising initial results. Our work has important implications for both research and clinical practice. Manual coding of clinical interactions using traditional qualitative coding techniques is a resource-intensive process, requiring a large investment of manpower. In addition, using human coders introduces the opportunity for bias associated with human subjectivity and coder fatigue. Utilizing automated coding algorithms has the potential to eliminate the bias as well as increase efficiency. For clinicians, this means more timely access to information that can directly inform their clinical practice and the care they provide to specific patients. As a future work, we are planning to experiment with additional features and more sophisticated classification models to further improve classification accuracy for all tasks.

6. REFERENCES

[1] S. Bergsma, M. Post, and D. Yarowsky. Stylometric analysis of scientific articles. In *Proceedings of NAACL-HLT'12*, pages 327–337, 2012.

[2] O. de Vel, M. Corney, A. Anderson, and G. Mohay. Language and gender author cohort analysis of e-mail for computer forensics. In *Proceedings of Digital Forensic Research Workshop*, 2002.

[3] R. Lacson and R. Barzilay. Automatic processing of spoken dialogue in the home hemodialysis domain. In *Proceedings of the 2005 AMIA Annual Symposium*, pages 420–424, 2005.

[4] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of NAACL-HLT'11*, pages 142–150, 2011.

[5] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *Proceedings of AAAI Workshop on Learning for Text Categorization*, 1998.

[6] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of EMNLP'02*, pages 79–86, 2002.

[7] S. Wang and C. D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of ACL'12*, pages 90–94, 2012.