Comparing Frequency- and Style-Based Features for Twitter Author Identification

Rachel M. Green
Department of Computer Science
The Johns Hopkins University
Baltimore, MD 21218
rgreen45@jhu.edu

John W. Sheppard
Department of Computer Science
Montana State University
Bozeman, MT 59717
john.sheppard@cs.montana.edu

Abstract

Author identification is a subfield of Natural Language Processing (NLP) that uses machine learning techniques to identify the author of a text. Most previous research focused on long texts with the assumption that a minimum text length threshold exists under which author identification would no longer be effective. This paper examines author identification in short texts far below this threshold, focusing on messages retrieved from Twitter (maximum length: 140 characters) to determine the most effective feature set for author identification. Both Bag-of-Words (BOW) and Style Marker feature sets were extracted and evaluated through a series of 15 experiments involving up to 12 authors with large and small dataset sizes. Support Vector Machines (SVM) were used for all experiments. Our results achieve classification accuracies approaching that of longer texts, even for small dataset sizes of 60 training instances per author. Style Marker feature sets were found to be significantly more useful than BOW feature sets as well as orders of magnitude faster, and are therefore suggested for potential applications in future research.

Introduction

Author identification is a subfield in Natural Language Processing (NLP) that uses machine learning techniques to determine the author of a text based on identifying characteristics such as word frequency, vocabulary, etc. (Manning and Schütze 1999). The majority of research in this field has focused on long formal texts such as excerpts from novels. However, as the current trends in information technology encourage an abundance of short, informal writing, it becomes increasingly important to determine to what extent author identification techniques also apply to short, informal text. Examples of such sources include email, forums and messaging boards, blogs, social networking sites such as Facebook, text messages, etc. There are many potential practical applications for author identification in these areas, such as identifying the source of anonymous messages for security purposes or using identification as the basis for developing targeted advertising. This paper focuses on short texts retrieved from Twitter (www.twitter.com), a social networking site that limits users to 140 character messages, commonly referred to as "tweets." We examine potential avenues of author identification in Twitter using supervised learning methods for data classification. Specifically, experiments were conducted using Support Vector Machines (SVM) with a variety of feature set options.

SVMs are an attractive approach to classification because they are designed to handle high-dimensional data, and have applied successfully to author identification in previous research (Diederich et al. 2000; Corney, A., and de Vel 2001; Stamatatos 2009; Inches and Crestani 2011). The extremely short length of Twitter messages presents a rarely examined challenge. It has long been assumed that there is a lower limit for text length under which identifying characteristics could no longer be apparent. Ledger and Merriam (1994) estimated this minimum length to be 500 words, and Forsyth and Holmes (1996) lowered it to 250 words based on a review of texts used in related studies. While this number does not reflect the abundant advances in statistical analysis and machine learning techniques since 1996, it is at least an order of magnitude higher than the length of Twitter messages; tweets average less than 25 words (often less than 10). While author identification has recently been shown to be possible in Twitter messages, results were limited (Sousa et al. 2011). We intend to demonstrate that author identification in this medium can approach accuracy levels closer to that of long text, which has been shown to have classification accuracy as high as 98% (Uzuner and Katz 2005).

For this paper, we conducted a series of 15 experiments designed to determine which type of feature set is most effective. The feature sets tested include Bag-of-Words (BOW) and Style Markers. Both feature sets were extracted from the same raw data. We hypothesized that the Style Markers feature set would perform better than the BOW feature set in terms of accuracy, as we believed the length of Twitter messages would not allow authors to express a large enough vocabulary for BOW to be effective.

Related Work

The problem of author identification via automated methods has been studied for over half a century, starting when Mosteller and Wallace (1964) applied a Bayesian statistical analysis method to determine authorship of "The Federalist Papers." While the quantity of work performed in the field

since then is too great to review in detail, Stamatos (2009) provides a good overview and comparison of author identification methods. Uzuner and Katz (2005) provide another.

The majority of research focused on large, formal text rather than formats similar to Twitter. Comparisons of BOW to Style Markers tended to show better performance with BOW in this format (Kaster, Siersdorfer, and Weikum 2005). However there is some research into shorter texts, sometimes utilizing Style Markers. Author identification in email is most studied, primarily through the use of data mining and SVM learning (de Vel 2000; de Vel, A., and G. 2001; Corney, A., and de Vel 2001). Zheng *et al.* (2006) examine and develop a framework of features that are likely to be useful for author identification in general online messages such as forum postings. Mohtasseb and Ahmed (2009) examine author identification specifically in blogs, comparing the effectiveness of Naïve Bayes (NB) and SVM algorithms. Fissette wrote her thesis on author identification in Dutch bulletin board messages, also making use of SVMs (Fissette 2010). In all cases, when researchers compared Style Markers and BOW, Style Markers outperformed BOW for short, informal text.

A small amount of existing research relates directly to Twitter. Lake (2010) discusses NLP in Twitter for purposes of data extraction. Author identification is not examined in this paper, but relevant issues such as data structure are addressed. Inches and Crestani (2011) also discuss Twitter in terms of data mining in text. Dietrick et al. (2012) examine gender identification in Twitter and Bergsma et al. (2012) examine automatic language identification. Author identification in Twitter is examined by Sousa et al. (2011), but their research is limited to 3 authors. Our research significantly expands the current knowledge by examining up to 12 authors, expanded feature sets and additional classifiers.

Experimental Approach

We developed a framework for experimentation designed to meet the goal outlined in the introduction. Our approach consists of the following four steps: Collect Data, Extract Features, Build Models, Evaluate Results.

Step 1: Data was retrieved from Twitter for several authors in a single domain. Due to the enormous number of authors on Twitter, five criteria were defined as a basis for author selection to narrow the scope while ensuring optimal data availability and minimizing bias.

1. Frequent new messages (average 20+ tweets/day)
2. Primarily new material (as opposed to re-tweets or quotes from other sources)
3. Individual rather than corporate or group authorship
4. Relatively consistent style or voice preferred (however this is subjective and hard to control)
5. Same domain or subject matter focus across all authors

Twelve authors were identified that meet these criteria, all from the financial field. Authors from only one domain were chosen to reduce unintentional bias introduced via the domain effect that could lead to trivial identification, particularly in the BOW model. The 12 selected authors are defined in Table 1 along with the amount of raw data collected after preprocessing. Due to limitations with the Twitter API, the amount of data available for each author is unbalanced. Experiments are based on equal sized randomly selected subsets of the total retrieved data.

Author	Twitter Handle	Instances
Jim Cramer	@jimcramer	1169
Lauren Young	@laurenyoung	986
Suze Orman	@suzeormanshow	169
Henry Blodget	@hblodget	463
Melinda Emerson	@smallbizwoman	214
Josh Brown	@reformedbroker	910
Tadas Viskanta	@abnormalreturns	270
Chris Adams	@chrisadamsmkts	431
John Carney	@carney	294
Ben White	@morningmoneyben	145
Gary Vaynerchuk	@garyvee	569
Jerry Khachoyan	@thearmotrader	373
Collection statistics: Average number of words per tweet (for all authors): 12.5 Average total characters per tweet (for all authors): 81.2		

Table 1: Twitter Authors Used

Raw datasets were preprocessed to remove messages expected to provide no value to the learning and testing processes. These were defined as messages containing less than 3 words and messages containing "re-tweets." Other forms of preprocessing commonly used with formal text were considered but ultimately deemed unnecessary for these datasets. For example, a stop list would remove a sizable percentage of words from many messages since tweets contain so few words in total. This may even be detrimental to author identification, since these words may otherwise represent a measurable feature.

Step 2: The data gathered and preprocessed in Step 1 was then converted into usable datasets by extracting the feature sets required for each experiment. For Feature Set 1 (BOW), all words found in the full dataset were considered as an unordered collection. Each significant word (defined as a word with a term frequency greater than five across the entire dataset of that experiment) was used as a feature. This resulted in large, sparse datasets with hundreds of features. For Feature Set 2 (Style Markers), 86 features were extracted based on style information. Extracted features included counts of various units within each message (e.g., characters, long words, whitespace, punctuation, hyperlinks, parts of speech), overall message characteristics (e.g., total length, total words), and frequency values (e.g., characters per message by %, punctuation per message by %).[1]

Step 3: Because of the focus of these experiments is on novel features rather than algorithms, we used the Sequential Minimal Optimization (SMO) algorithm included in WEKA 3.6.8 (Hall et al. 2009) [2]. A linear kernel with default param-

[1] A complete list of the features can be found in the technical report version of this paper at http://nisl.cs.montana.edu/~nisldata/.

[2] Software available for download at http://www.cs.waikato.ac.nz/~ml/weka/

eters was used. We performed each experiment using a 5x2 cross-validation design, as proposed by Dietterich (1998).

Step 4: A series of experiments was conducted with each feature set (15 experiments total), evaluated primarily for classification accuracy. Each experiment was run multiple times with values such as root mean squared error (RMSE) recorded to verify statistical significance. Build time was also noted as a measure of efficiency.

Results

A total of 15 experiments were conducted. Each experiment followed a 5x2 cross validation scheme based on a different random seed. Conducting five 2-fold runs of each experiment was intended to minimize anomalous results and demonstrate statistical significance. Values shown in Table 2 correspond to the average of all runs for each experiment. Table 3 shows the confusion matrix with percent classified in each category. The full experimental results are available in an online Appendix. [3]

The following naming convention is used for Experiments in Table 2: ⟨feature set⟩⟨number of authors⟩: ⟨total instances per author⟩. Feature sets can be BOW (Bag-of-Words) or SM (Style Markers).

The first 4 experiments were designed to determine which of the two feature sets performed better. Parallel experiments were conducted using each feature set extracted from the same dataset. The Style Markers feature set was found to be much more effective than the BOW feature set in all experiments. In the parallel experiments 1 and 3, the average classification accuracy jumped from 76.7% for BOW to 92.3% for SM. Considering the BOW model took over 25 times longer to build, the Style Markers feature set is a clear winner. This is supported by the parallel experiments 2 and 4, in which classification accuracy from BOW to SM rose from under 60% to over 75% with a 62-fold decrease in training time. Additional BOW experiments were not conducted based on these results, as Style Markers were already shown to be more effective.

Experiments 5–15 were conducted to determine at what rate the classification accuracy of the Style Markers feature set diminished as more authors and smaller datasets were used. Experiment 5 began by dropping the dataset size from 900 to 120 instances per author. Subsets were chosen at random from the available data with the same random subset for each author used throughout experimentation. We found that using this much smaller dataset had no discernible effect on classification accuracy. Furthermore, the drop in dataset size led to a considerable decrease in build time. To limit bias as much as possible, the dataset used for Experiment 5 was a subset of the dataset used in Experiment 3. Experiment 6 parallels Experiment 4 in the same way. Using a subset of the same data for 3 authors, classification accuracy actually rose from 74.9% to 83.0% when the number of instances was dropped from 900 to 120 per author, although this rise may not be significant.

In each subsequent experiment, data from an additional author chosen at random was added to the current set. As authors increased, the classification accuracy dropped quickly until reaching 5 authors, after which it continued to drop at a much slower rate (Figure 1). Even using the full set of 12 authors with only 120 instances each (therefore 60 training instances), classification accuracy remained above 40%. This is considerably higher than anticipated. Possible explanations are discussed in the next section.

[3] http://nisl.cs.montana.edu/~nisldata/

Experiment	Accuracy	RMSE	Build Time (sec)
BOW2: 900	76.67485	0.48285	44.985
BOW3: 900	59.99455	0.434375	395.4425
SM2: 900	92.3203	0.27665	1.7575
SM3: 900	75.108925	0.383	6.3825
SM2: 120	92.378025	0.27545	0.0975
SM3: 120	80.737675	0.353775	0.1725
SM4: 120	70.398775	0.362825	1.1175
SM5: 120	58.333325	0.356075	0.615
SM6: 120	51.988275	0.328075	2.54
SM7: 120	53.33334	.32048	0.9
SM8: 120	47.29166	0.30844	1.02
SM9: 120	46.79835	0.2945	3.0575
SM10: 120	44.01668	0.2834	1.71
SM11: 120	43.45454	0.27316	1.964
SM12: 120	40.541925	0.26425	2.485

Table 2: Experimental Results

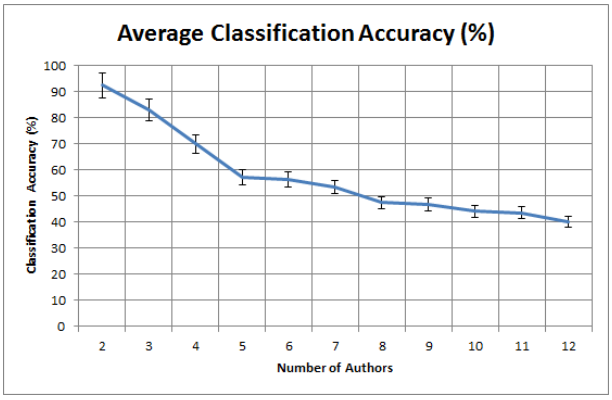


Figure 1: Average classification accuracy (%) for multiple authors (based on sets of 120 instances each): Actual values shown with 95% confidence interval.

Discussion

As hypothesized, Style Markers were more effective for classification than BOW. There are many possible explanations for this.

- The limiting factor of 140 characters per Twitter message ensures relatively few total words. This may limit the scope of word-frequency based methods including BOW.
- The 140 character per message limitation may encourage authors to choose their words carefully or to use uncharacteristic words to meet the space requirement, thus limiting individualized vocabulary. Again, this may limit the effectiveness of BOW.

Classify as ⟹	a	b	c	d	e	f	g	h	i	j	k	l
a. Tadas Viskanta	0.42	0.08	0.00	0.03	0.09	0.07	0.09	0.05	0.07	0.03	0.05	0.01
b. Jim Carney	0.11	0.21	0.06	0.01	0.15	0.05	0.14	0.09	0.06	0.02	0.09	0.03
c. Chris Adams	0.05	0.06	0.55	0.03	0.03	0.02	0.09	0.02	0.03	0.02	0.07	0.04
d. Gary Vaynerchuck	0.02	0.00	0.00	0.66	0.03	0.12	0.00	0.01	0.01	0.15	0.00	0.01
e. Ben White	0.11	0.12	0.05	0.05	0.22	0.16	0.04	0.05	0.03	0.08	0.06	0.02
f. Suze Orman	0.03	0.02	0.01	0.12	0.10	0.62	0.01	0.02	0.01	0.04	0.00	0.02
g. Lauren Young	0.04	0.10	0.05	0.01	0.02	0.02	0.55	0.02	0.01	0.03	0.13	0.03
h. Jim Cramer	0.09	0.12	0.04	0.01	0.09	0.07	0.07	0.31	0.11	0.02	0.04	0.04
i. Jerry Khachoyan	0.16	0.07	0.04	0.06	0.11	0.07	0.03	0.14	0.18	0.08	0.06	0.01
j. Josh Brown	0.04	0.01	0.01	0.35	0.07	0.14	0.01	0.04	0.04	0.27	0.00	0.03
k. Henry Blodget	0.04	0.09	0.04	0.00	0.08	0.01	0.20	0.05	0.04	0.01	0.39	0.06
l. Melinda Emerson	0.03	0.04	0.02	0.10	0.05	0.07	0.07	0.02	0.03	0.06	0.07	0.44

Table 3: Confusion Matrix—Twelve Authors

- Many universal acronyms and abbreviations have become popular in Twitter due to the short length of messages (such as CX for ”correction” or AFAIK for ”as far as I know”). Authors may be using these instead of their own characteristic word choices.
- Compared to word choice, the 140 character limitation may not have as pronounced an effect on style markers such as use of capitalization and punctuation, which are not significantly hindered by message length
- Style markers may be more effective at addressing data sparsity than BOW.

Both feature sets were far more effective than previously expected. Classification accuracy reached 92.2% for two authors and remained above 40% even when identifying among a dozen different authors using sets of only 120 tweets from each. This is about five times more accurate than random chance, even without any optimization or modification of the standard algorithm in use. The following may contribute to these unexpected results:

- The informal text style of Twitter allows individuality that would not be apparent in formal text, such as the use of slang and departure from formal grammar. These features may be easier to detect in small samples than formal characteristics.
- The short length of tweets may actual encourage authors to adopt stylistic characteristics they would not use otherwise, such as omitting articles or using characteristic abbreviations (such as substituting ”U” for ”you”). Preliminary analysis of the raw data suggests that most authors remain fairly consistent in which methods they adopt to stay within space limitations.
- Because tweets imitate spoken language, voice and mood play a more dominant role than in formal text. Authors may adopt different methods of showing emphasis or conveying mood. For example, some authors may use words in all capital letters to express importance whereas others may prefer exclamation points, emoticons, surrounding asterisks or other methods. Again, preliminary analysis suggests that authors are fairly consistent in their preferred method.
- Authors appear to have varying primary modes for using Twitter, even within the domain-specific group selected for these experiments. For example, some authors appear to focus on Twitter as a social medium, with emphasis on referencing and responding to other users directly, or posting personal status or location updates. Other authors appear to use Twitter as a short-format blog, with the majority of their messages focusing on imparting general opinions and information. The mode of use may influence content choice and general stylistic markers such as frequency of user references, hashtags, hyperlinks, etc.

Some stylistic characteristics are immediately apparent in the raw data, as shown in Table 4, even when users post similar content. For example, both @jimcramer and @henryblodget refer directly to Apple stock, however, @jimcramer uses the stock ticker reference “$AAPL” whereas @henryblodget spells out the entire word “Apple.” Note also @jimcramer’s use of exclamation points versus @henryblodget’s use of capitalized words for emphasis.

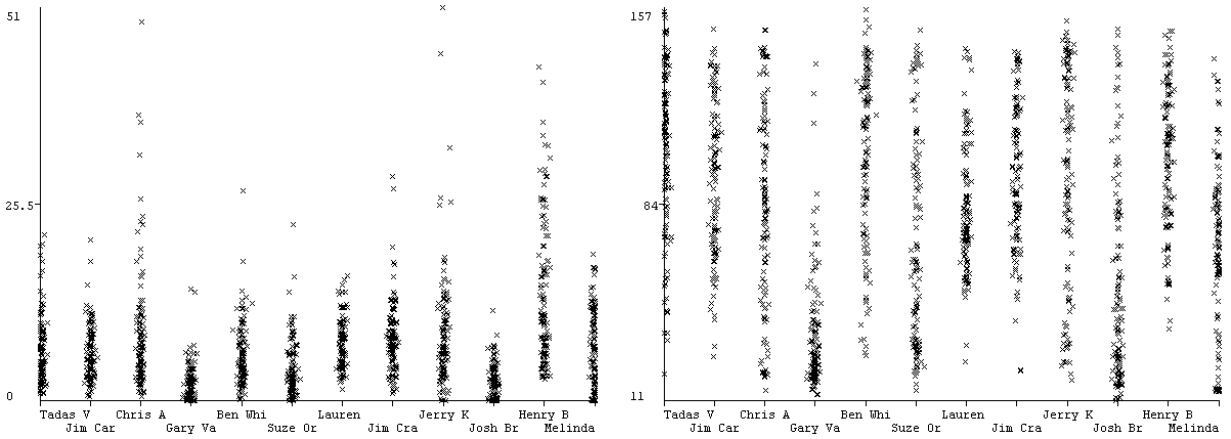
These differences lead to identifiable patterns characteristic to each author for each type of style marker in the feature set. Although there is a good deal of overlap between authors, the combined result of multiple style markers may represent a “fingerprint” for each individual that can be identified by machine learning techniques. Figures 2(a)–2(b) show a few examples from the data used in these experiments. More examples of these patterns can be seen in the technical report. Some style markers appear to be more effective than others in differentiating various authors. Future research may indicate that a smaller subset of the total 86 features examined in this research effectively determine the accuracy of author identification. In this case, the smaller subset of features could be used, thus further reducing the feature set size and total processing time.

Examining the confusion matrix for the 12 author case may help explain some of these observations.

- The percentage of correct classification for each author ranged from 18% to 66% for the 12 author case (2.2 – 8x random chance). This implies that there may be a wide variation in effectiveness of author identification for individuals based on personal style or other factors.

Author	Sample Tweets
@jimcramer	$LNG gets approval for export terminal.. Here comes the jobs! There goes the gas....
	Countdown in 4 minutes–7th Anniversary Mad Money!!!!!!!!!!!!!!!!!!!
	Fired up discussion coming on $AAPL tear down, thanks for all of your help on this. Just got sodabread–1000 crunches coming!!!
@henryblodget	Buying lottery tickets and playing blackjack, meanwhile (both legal) are not risky. They're GUARANTEED LOSSES. Gov't allows this
	CALLING ALL MUPPETS: Apple just blew through $600... $500 or $700 next? VOTE NOW! http://t.co/Fi52gjj1
	Hey, @mashable, how are those CNN negotiations going?

Table 4: Sample Data



(a) Pattern of usage for capital letters

(b) Pattern of usage for total characters

Figure 2: Patterns of usage

- The correct classification was not the highest classification in only one case (@reformedbroker). This confusion was not mutual; in fact the other author (@garyvee) demonstrated the highest classification accuracy of the 12. The reason for this discrepency is not clear, however the data shows that the two authors have similar averages for some key features such as "user reference" (1.233 vs 1.225). This may imply that certain features are more important than others for correct classification.
- Some differences are immediately apparent in the original raw data for the best and worst authors. The highest performing author (@garyvee) began almost every tweet with a reference or reply to another user. The lowest performing author (@thearmotrader) used a wide variety of message types and formats. It is likely that this relative inconsistency contributed to the classification confusion.
- Surprisingly, the average word count for the best author was almost three times lower than average word count for the worst author at 5.5 words vs. 15 words. This implies that stylistic features do not require more than a minimal length to be effective.

Future Work

The results of this paper suggest several avenues of future research.

First, to determine whether our results are valid in the general case, additional experiments should be performed using different data. This would reduce the possibility that a bias introduced by author selection influenced the results. It would also be interesting to expand experimentation to determine the rate at which classification accuracy falls in a larger sample size. While the results with twelve authors are promising, real-world applications may require discrimination between hundreds or thousands of authors to be useful. Additionally, it may be worthwhile to investigate how these results apply to a more diverse or natural group of authors as opposed to authors selected exclusively from one domain. Since the Style Markers feature set was determined to be more effective than the BOW model, the domain effect based on content may not be an issue. This may also help determine whether an unintentional bias was introduced by the specific set of raw data gathered for these experiments.

Since the Style Markers feature set was determined to be more effective, additional research in this direction is suggested, for example by including additional features such as vocabulary richness. It may also be possible to incorporate metadata such as message timestamp as a feature, as there may be detectable patterns in the time and frequency of postings. Re-tweets may also provide another valuable feature. Although they were removed completely from the datasets

in these experiments, we noted that most authors precede a re-tweet with a statement of their own. It may be more effective in the future to remove only content following the re-tweet flag instead of the entire message, which would make frequency of re-tweets another measurable feature.

Alternatively, it may make more sense to reduce the number of features rather than add more. Additional analysis could narrow the scope of the feature set by determining which features provide the best results. This would reduce the size of each data point and speed the learning process, which may be important if considerably larger datasets are used. Alternative feature extraction techniques could also be explored, such as latent topic models (e.g., latent semantic analysis).

Furthermore, it would be interesting to research whether the superior effectiveness of style markers over other methods would also apply to text formats beyond Twitter.

Conclusion

The experimental results presented in this paper suggest that the conventional assumption of a minimum threshold for author identification does not apply to style marker feature sets. The average text length of tweets used in these experiments was 12.5 words (81.2 characters), and classification accuracies on these datasets were consistently far above random chance, even with small set sizes. The results of these experiments suggest that author identification via style marker feature sets may be more effective than traditional methods of semantic or word analysis for short text.

References

Bergsma, S.; McNamee, P.; Bagdouri, M.; Fink, C.; and Wilson, T. 2012. Language identification for creating language-specific twitter collections. *NAACL-HLT 2012* 65.

Corney, M., A.; A., Mohay, G.; and de Vel, O. 2001. Identifying the authors of suspect e-mail. *Computers and Security (submitted)*.

de Vel, O., A.; A., Corney, M.; and G., M. 2001. Mining e-mail content for author identification forensics. *SIGMOD* Rec. 30(4):55–64.

de Vel, O. 2000. Mining e-mail authorship. In *Proceedings of the Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD 2000)*.

Diederich, J.; Kindermann, J.; Leopold, E.; and Paass, G. 2000. Authorship attribution with support vector machines. *Applied Intelligence* 19:109–123.

Dietrick, W.; Miller, Z.; Valyou, B.; Dickinson, B.; Munson, T.; Hu, W.; Garzia, F.; Tirocchi, N.; Scarpiniti, M.; Cusani, R.; et al. 2012. Gender identification on twitter using the modified balanced winnow. *Communications and Network* 4(3):189–195.

Dietterich, T. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* 10:1895–1923.

Fissette, M. 2010. Author identification in short texts. Bachelor's Thesis, Department of Artificial INtelligence, Radboud University.

Forsyth, R., and Holmes, D. 1996. Feature finding for text classification. *Literary and Linguistic Computing* 11(4):163–174.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.* 11(1):10–18.

Inches, G., and Crestani, F. 2011. Online conversation mining for author characterization and topic identification. In *Proceedings of the 4th Workshop for Ph.D. Students in Information and Knowledge Management*, 19–26. New York, NY: Association of Computing Machinery.

Kaster, A.; Siersdorfer, S.; and Weikum, G. 2005. Combining text and linguistic document representations for authorship attribution. In *SIGIR workshop: Stylistic Analysis of Text For Information Access*.

Lake, T. 2010. Status report: Twitter nlp. Western Michigan University, Kalamazoo.

Ledger, G., and Merriam, T. 1994. Shakespeare, fletcher, and the two noble kinsmen. *Literary and Linguistic Computing* 9:235–248.

Manning, C., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mohtasseb, H., and Ahmed, A. 2009. Mining online diaries for blogger identification. In *Proceedings of the International Conference of Data Mining and Knowledge Engineering – The World Congress on Engineering*.

Mosteller, F., and Wallace, D. 1964. *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.

Sousa, R.; Laboreiro, G.; Sarmento, L.; Grant, T.; Oliveira, E.; and Maia, B. 2011. Automatic authorship analysis of micro-blogging messages. *Natural Language Processing and Information Systems* 161–168.

Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60(3):538–556.

Uzuner, O., and Katz, B. 2005. A comparative study of language models for book and author recognition. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, 969–980. Berlin, Germany: Spring.

Zheng, R.; Li, J.; Chen, H.; and Huang, Z. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society of Information Sciences* 57:378–393.