SEOULTECH

Authorship Attribution for Social Media Forensics

IEEE TRANSACTIONS ON INFORMATION FORENSISCS AND SECURITY, Jan, 2017

Anderson Rocha, Christopher W. Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne R. B. Carvalho, and Efstathios Stamatatos

2017/2 - Digital forensics

17510140 신국희
17512084 마상열

CIEL Communication & Intelligence Laboratory 통신및지능연구실

https://sites.google.com/site/cielseoultech/

CONTENTS

1. Introduction
2. A review of method in authorship attribution relevant to social media forensics
3. Walk-through of authorship attribution techniques for social media forensics
4. Experimental results
5. Conclusion and future directions

Abstract

- The veil of anonymity provided and distributed networks has drastically complicated the task of identifying users of social media during forensic investigations.
- In some cases, the text of a single posted message will be the only clue to an author's identity.
- For the past 50 years, linguists, computer scientists, and scholars of the humanities have been jointly developing automated methods to identify authors based on the style of their writing.
- All authors possess peculiarities of habit that influence the form and content of their written works.

In this paper, provide a comprehensive review of the methods of authorship attribution.

Introduction

A. **Motivation for automated authorship attribution methods for social media forensics**

B. **Contributions of this review article**

- It is well known that the real lives of Internet users sometimes turn out to be entirely different from who they appear to be online, but this phenomenon are changing.
- Russian media agency that allegedly executed organized disinformation campaigns on social media using pseudonyms and virtual identities.
- The agency achieved success in promoting false news events and influencing public opinion on politics, and was even able to deceive the journalist covering the story for the Times.
- On the Internet, It is called “trolling”, this poses a legal and security dilemma on multiple fronts.

I.A : Motivation for automated authorship attribution methods for social media forensics

- The enabling factor in "trolling" is a reliance on anonymity to ensure the success of a social media campaign. (ex pre-paid SIM, unlocked smartphone)
- Also, Many users are turning to encrypts user data and randomly sends through various nodes for anonymization.(Using Tor service and Onion Routing network)
- one can simply tunnel traffic through a series of proxy servers, many of which are freely available and open to the public.
- As a result, anonymity can frustrate an investigation to the point where network forensics cannot be used.
- The text left on a social media platform may be our only clue to the author's identity.

I.A : Motivation for automated authorship attribution methods for social media forensics

- The goal of Authorship Attribution is to identify authors of texts through features derived from the style of their writing. (It is called Forensic Authorship Attribution)
- In contrast to other authorship attribution tasks found in the broader world, forensic authorship attribution does not assume a claimed identity before analysis to verify a match.
- Instead, it is assumed that the source of a text is either one author out of a known set, or an author that is unknown to investigators.
- Automated approaches to authorship attribution via the methods of statistical pattern recognition have been around for decades, with more recent work evaluating the utility of advanced machine learning techniques.

I.A : Motivation for automated authorship attribution methods for social media forensics

- Authorship attribution as an academic discipline has maintained a unique relationship to scholarship in the digital humanities, where the authentication of disputed literary works is of interest.
- Much of the existing work in this area remains unfamiliar to practitioners and researchers in forensics.
- Relationship between the work of humanists and forensic examiners cannot be denied.
- Notwithstanding, the underlying problem domains humanists and forensic examiners operate in can be rather different.
- In response to this, some researchers suggest joining the messages in a single document.

I.A : Motivation for automated authorship attribution methods for social media forensics

- we know that there is often enough distinct information in even just a handful of sentences for a human reader to understand that they are from a common source.

A.1: A beautiful reflection on mortality by a great man.
A.2: Unintentional reductio ad absurdum: "Colleges Need Speech Codes Because Their Students Are Still Children"
A.3: The great taboo in discussions of economic inequality: Permanent income has nonzero heritability.

And these from Author B:

B.1: Incredible wknd w/ @CASAofOC. Thx to all of the folks that helped raise $2.8 million to help young people in need.
B.2: Thx 4 having me. Great time w/ you all
B.3: Great to meet you & thx for your support @ChrissieM10

- A : tweets tend to be well composed and include diverse vocabulary ("mortality," "reductio ad absurdum," "taboo," "heritability")
- B : including frequent use of abbreviation ("Thx," "wknd").

I.A : Motivation for automated authorship attribution methods for social media forensics

- To adapt authorship attribution to social media, we need to find stylometric features that capture the diversity of the language deployed therein.
- Given that the messages are short, this task requires a large amount of training data to increase the classification accuracy.
- It must be emphasized that even the most promising techniques in authorship attribution are not nearly as precise as DNA testing, and forensic stylometry will rarely, if ever, be used in a courtroom by itself.
- The most promising techniques in authorship attribution are not nearly as precise as DNA testing, therefore forensic stylometry will rarely, if ever, be used in a courtroom by itself.

I.B : Contributions of This Review Article

- An overview of forensic authorship attribution, with a discussion of why it is distinct from more general authorship attribution applications found in audio, speech and language processing.
- A comprehensive review of the methods for authorship attribution that are applicable to forensics in a social media context.
- An analysis of feature sets that can extract a large amount of information from a small amount of text, which is crucial for achieving good performance in a forensic setting.
- A detailed walk-through of supervised learning methods for authorship attribution for social media forensics.
- A discussion of open problems in forensic authorship attribution.

critical review of existing methods for authorship attribution and how they relate to forensic authorship attribution and social media

A review of methods in authorship attribution relevant to social media forensics

A. General stylometric Features for Forensic Authorship Attribution
B. General Classifiers for Forensic Authorship Attribution
C. Authorship Attribution for Small Samples of Text
D. Authorship Attribution Specifically for social Media
E. The Threat of Counter-Attribution

- forensic authorship attribution is different from more general authorship attribution applications found in audio, speech and language processing.
- One such general attribution application is Authorship Verification.
- Authorship verification is a 1:1 classification task with a known positive class, and a negative class of "all texts by all other authors" that is vast and extremely heterogeneous.
- We more often are faced with a problem of 1:N identification
- Unlike authorship in the digital humanities, the question of how to falsify a prediction is important for the admissibility of evidence.
- In all cases of forensic authorship attribution, how we treat the task is instrumental to the design of appropriate algorithms.

- The key considerations for forensic authorship attribution are :
  - No control over the testing set that predictions will be made from which could be limited to a single sample.
  - No control over the quality of the training data used to create the attribution classifiers.
  - The determination of a well-defined error rate for an algorithm, before it is applied to a real-world problem.
  - It is possible that the author under investigation is deliberately evading automated attribution methods.

- The components we need to build a computational pipeline that is suitable for forensic authorship attribution for social media analysis are:
  - A training corpus drawn from postings to social media.
  - A pre-processing regime that filters and extracts features from text.
  - A classifier training regime that yields author-specific models.
  - A decision making process that makes a prediction for an anonymous tweet based on the classifier at hand, after the source data has been pre-processed and transformed into a bag-of-words representation.

- At the most basic level, the words of a text are useful features for authorship attribution.
- However, all words cannot simply be treated as features.
- It is common to discard the function words, those words that occur most frequently
- Term Frequency-Inverse Document Frequency(TF-IDF)
  - It is a weighted measure of word importance that is commonly used in information retrieval tasks.
  - It is calculated by multiplying the term frequency by the inverse document frequency for a specific word.
  - It emphasizes the importance of key words commonly deployed by a specific author, while deemphasizing those that are function words.
- Beyond word rarity measures, more elemental probabilistic features of language prove to be very effective for attribution tasks.

❑ Word-level N-gram

- It is a feature that represents the probability of an element e occurring given some history h, or P(e|h).
- The advantage of using n-grams is that they capture lexical preferences without the need of any *a priori* knowledge of a grammar for a language, which is necessary to identify features like function words.
- feature is that some authors might have a preference for some expressions composed of two or more words in sequence, the probability of which is captured by n-grams of these words.

❑ "Stamatatos"

- studied the robustness of n-grams for authorship attribution.
- shown that the vast majority of existing research in this area only examines a case in which the training and testing corpora are similar in terms of genre, topic, and distribution of texts.

- "Sapkota"
  - show that not all n-grams are created equal, and group them into three categories: morphosyntax, thematic content and style.
  - categorization improves our understanding of authorship attribution for a social network combining numerous demographics and special interests.
- Natural Language Processing (NLP) toolkits
  - The availability of natural language processing (NLP) toolkits for many languages enables the use of more complicated stylometric features based on syntactic or semantic analysis of texts.
  - measures are noisy and less effective than n-grams.

❑ Part-of-Speech(POS) tagging

- It is a potentially rich source of additional feature information.
- The motivation is that grammatical usage can serve as an important indicator of an author's style even in short messages.
- The features may be effective in isolation, or they may be used to augment other lexical feature sets.
- Such an approach can also incorporate Twitter-specific content because hashtags, links, retweets, etc. are assigned to specific POS tags.
- POS tagging is a standard tool in NLP, we suggest for the first time its use coupled with supervised machine learning in authorship attribution for social media

"A short sentence."			
'A'	'short'	'sentence'	'.'
Indefinite Article	Adjective	Noun	Punctuation

- Feature set has been chosen, the next step is to select a classification method.
- Unsupervised clustering is appealing, in that there is no need to assemble large pre-labeled training sets before making use of an attribution algorithm.
  - The algorithm generates vectors of frequencies of function words and applies Principal Component Analysis (PCA) over them. Authors are then classified via data clustering.
- Multivariate analysis achieved some measure of success, and it quickly became well-established for authorship attribution.
- However, the presence of some labeled data often improves results dramatically. Accordingly, supervised approaches now dominate the field.

- Simple supervised classification methods are often dismissed in favor of the more elaborate, highly parametrized, algorithms that are prevalent in the current literature.
- of course, should not be ignored in cases where a desired error rate can be achieved without a large amount of computational time or extensive parameter tuning.
- For attribution problems, this means formulating a null hypothesis that suggests that two works under consideration are from different authors, and testing it via a hypothesis test(e.g., Student's t-test).
- Of the various hypothesis tests that have been proposed, $\chi 2$ is particularly attractive in that it can yield a ranked list of candidate authors in a 1:N attribution scenario.

- Another straightforward classification method is to compute distance to known vectors, and assign authorship based on the shortest distance to a known vector that passes a particular matching threshold.
- With the availability of good quality labeled data, supervised classifiers can be trained to make use of common patterns across many samples, instead of making direct comparisons between individual samples.

- K-nearest Neighbors (K-NN)
  - K-nearest Neighbors (K-NN) assigns class membership based on the distance of an unknown point to K points that are closest to it.
  - if the majority of those K nearest points are from the same author, we can conclude that the unknown sample should also be associated with the author of those points.
- Another supervised classification method is Naïve Bayes.
  - When applying Naïve Bayes to authorship attribution, the resulting probability value for class assignment can be consulted to determine if a match to a known author has been made.
- Markov Models for authorship attribution calculate probabilities of letter or word transitions, which are style-specific markers, placing them into author specific transition matrices.

- Kullback-Leibler Divergence is used to measure the relative entropy between the probability mass functions of features extracted from texts
  - authorship assigned to a pairing with the lowest relative entropy.
  - Highly Parameterized models are a better option to learn complex relationships in high dimensional feature spaces.
  - Multi-layer neural networks are well suited to learning models for non-linear problems, and can process an enormous diversity of feature sets.
  - While it is commonly alleged that it is not possible to determine the basis by which neural networks make classification decisions, the recent resurgence of this area has yielded work that shows that this is possible for some architectures in computer vision.

❑ Support Vector Machines (SVM)

- Following trends in supervised machine learning, Support Vector Machines (SVM) emerged as a method in authorship attribution to address the same problem of high-dimensional feature vectors
- In practice, the SVM's concept of maximum margin leads to better generalization, and thus better accuracy for binary and multi-class classification problems.
- they can easily overfit the training data when kernelized, and are more suited to binary problems.

❑ Decision Trees

- Use a graphical model over multiple inputs to assign a class label to incoming data, are an alternative.
- There is some utility to using decision trees on their own, but a meta-learning technique considering ensembles of them is more powerful.

❑ Random Forest

- Random Forest treats decision trees as weak learners, and randomly subsamples sets of features from a specific training dataset to improve accuracy and mitigate overfitting.
- Random forest classifiers are also attractive because they provide a clear indication of the feature weighting via variable importance measures.

TABLE VI
UNSUPERVISED CLASSIFICATION FOR AUTHORSHIP ATTRIBUTION

Source	Features Used	Classifier	Corpus
Burrows 1987 [28], 1989 [29], 1992 [30], [31]	Small set of function words	Multivariate analysis	English prose
Ledger and Merriam 1994 [117]	Character-level n-grams	Multivariate analysis	English drama
Mealand 1995 [132]	Function words	Multivariate analysis	Greek prose
Holmes and Forsyth 1995 [75]	Words	Multivariate analysis	English prose
Baayen et al. 1996 [13]	Syntax	Multivariate analysis	English prose
Merriam 1996 [137]	Function words	Multivariate analysis	English drama
Tweedie and Baayen 1998 [186]	Function words	Multivariate analysis	Latin prose
Binongo and Smith 1999 [20]	Function words	Multivariate analysis	English drama
Holmes et al. 2001 [76]	Function words	Multivariate analysis	Journalism
Baayen et al. 2002 [12]	Function words	Multivariate analysis	Dutch prose
Hoover 2003 [80]	Word-level n-grams	Multivariate analysis	English prose
Binongo 2003 [19]	Function words	Multivariate analysis	English prose
Kestemont et al. 2015 [99]	Function words	Multivariate analysis	Latin prose

II.B : General Classifiers for Forensic Authorship Attribution

TABLE VII
DISTANCE-BASED AND SIMPLE MODEL-BASED CLASSIFICATION FOR AUTHORSHIP ATTRIBUTION (SORTED BY CLASSIFICATION TYPE)

Bissel 1995 [22]	Weighted cum. sum of lexical statistics	Statistical hypothesis test	English prose
Somers 1998 [170]	Weighted cum. sum of lexical statistics	Statistical hypothesis test	English prose
Chaski 2001 [39]	Syntax and punctuation	Statistical hypothesis test	English prose
Somers and Tweedie 2003 [171]	Weighted cum. sum of lexical statistics	Statistical hypothesis test	English prose
Merriam 1979 [134], 1980 [135], 1982 [136]	Word positions	Statistical hypothesis test	English drama
Grieve 2007 [68]	Words, syntactic structures, and character-level n-grams	Statistical hypothesis test	English prose
Kjell 1994 [104]	Character-level n-grams	Cosine Similarity	English prose
Hoover 2004 [82]	Function words	Delta	English prose and poetry
Kestemont et al. 2015 [99]	Function words	Delta	Latin prose
Kukushkina et al. 2001 [114]	Character-level n-grams and grammatical word classes	Markov models	Russian prose
Khmelev and Tweedie 2002 [101]	Character-level n-grams	Markov models	English prose
Khmelev and Teahan 2003 [100]	Character-level n-grams	Markov models	English journalism
Zhao et al. 2006 [197]	Parts of speech	Kullback-Leibler Divergence	English novels and journalism
Zhao and Zobel 2007 [196]	Function words and part-of-speech tags	Kullback-Leibler Divergence	English prose and drama
Teahan and Harper 2003 [183]	Character streams	Cross-entropy	English journalism
Juola and Baayen 2005 [90]	Character streams and function words	Cross-entropy	Dutch prose
Kjell et al. 1995 [105]	Character-level n-grams	K-NN	English journalism
Hoorn et al. 1999 [78]	Character-level n-grams	K-NN and Naïve-Bayes	Dutch poetry
Keselj et al. 2003 [97]	Character-level n-grams	K-NN	English prose and Greek journalism
Zhao and Zobel 2005 [195]	Function words	K-NN and Naïve-Bayes	English journalism
Mosteller and Wallace 1964 [141]	Small set of function words	Naïve-Bayes	English prose
Clement and Sharp 2003 [42]	Character-level n-grams	Naïve-Bayes	English prose
Peng et al. 2004 [150]	Character- and word-level n-grams	Naïve-Bayes	Greek journalism
Savoy 2013 [161]	Function words	Naïve-Bayes	English prose
Stamatatos et al. 2000 [179], 2001 [180]	Syntactic chunks	Linear Discrim. Analysis	Greek journalism
Chaski 2005 [40]	Character- and word-level n-grams	Linear Discrim. Analysis	English prose
Jockers and Witten 2010 [87]	Words and word-level bigrams	Regularized Discrim. Analysis, Delta, and K-NN	English prose

II.B : General Classifiers for Forensic Authorship Attribution

TABLE VIII
MODEL-BASED CLASSIFICATION FOR AUTHORSHIP ATTRIBUTION (SORTED BY CLASSIFICATION TYPE)

Matthews and Merriam 1994 [129]	Small set of function words	Neural Networks	English drama
Merriam and Matthews 1994 [138]	Function words	Neural Networks	English drama
Kjell 1994 [102], [103]	Character-level n-grams	Neural Networks	English prose
Lowe and Matthews 1995 [120]	Function words	Neural Networks	English drama
Martindale and McKenzie 1995 [127]	Words	Neural Networks	English prose
Kjell et al. 1995 [105]	Character-level n-grams	Neural Networks	English journalism
Tweedie et al. 1996 [187]	Function words	Neural Networks	English prose
Hoorn et al. 1999 [78]	Character-level n-grams	Neural Networks	Dutch poetry
Waugh et al. 2000 [190]	Function words	Neural Networks	English prose
Zheng et al. 2006 [198]	Characters, function words and syntax	Decision Trees, Neural Networks, and SVM	English and Chinese newsgroups
Li et al. 2006 [118]	Lexical, syntactic, structural, and content-specific features	Neural Networks and SVM	English and Chinese newsgroups
Tearle et al. 2008 [184]	Lexical, syntactic, structural, and content-specific features	Neural Networks	English prose and English drama
Jockers et al. 2008 [88]	Words	Nearest Shrunken Centroid	English prose
Jockers and Witten 2010 [87]	Words and word-level bigrams	Nearest Shrunken Centroid and SVM	English prose
Schaalje and Fields 2011 [162]	Word-level statistics	Nearest Shrunken Centroid	English prose
Fung 2003 [63]	Function words	SVM	English prose
Diederich et al. 2003 [47]	Function words	SVM	German journalism
Gamon [64]	Function words, syntactic and semantic features	SVM	English prose
Koppel et al. 2005 [112]	Function words and part-of-speech tags	SVM	English prose
Koppel et al. 2006 [110]	tf-idf over words and characters	SVM	English web posts
Argamon et al. 2007 [9]	Functional lexical features	SVM	English prose
Pavelec et al. 2007 [148]	Conjunction types	SVM	Portuguese journalism
Koppel et al. 2007 [111]	Function words, syntactic structures, part-of-speech tags, complexity and richness measures, and syntactic and idiosyncratic usage	SVM	English prose

II.B : General Classifiers for Forensic Authorship Attribution

Stamatatos 2008 [174]	Character-level n-grams	SVM	English and Arabic journalism
Forstall and Scheirer 2009 [57]	Character-level n-grams	SVM	English prose, and English and Latin poetry
Escalante et al. 2011 [53]	Character-level n-grams	SVM	English journalism
de Vel et al. 2001 [46]	Capitalization, white space, and punctuation	SVM	English e-mail
Hedegaard et al. 2011 [73]	Word- and character-level n-grams and semantic features	SVM	English and (translated) Russian prose
Savoy 2013 [161]	Function words	SVM	English prose
Sidorov et al. 2014 [168]	Syntactic n-grams	SVM	English prose
Sapkota et al. 2015 [159]	Character-level n-grams	SVM	English electronic communication and journalism
Abbasi and Chen 2005 [1]	Lexical, syntactic, and structural features	Decision Trees and SVM	Arabic and English web posts
Argamon et al. [8]	Function words and part-of-speech tags	Decision Trees	English journalism
Koppel and Schler 2003 [106]	Function words, part-of-speech tags, idiosyncratic usage	Decision Trees and SVM	English e-mail
Popescu and Grozea 2012 [151]	Character-level n-grams	Random Forest	AAAC data set [89]
Bartoli et al. 2015 [15]	Lexical, syntactic, structural, and content-specific features	Random Forest	English, Dutch, Greek and Spanish prose
Maitra et al. 2015 [123]	Lexical, syntactic, structural, and content-specific features	Random Forest	English, Dutch, Greek and Spanish prose
Pacheco et al. 2015 [147]	Lexical, semantic, syntactic, structural, and content-specific features	Random Forest	English, Dutch, Greek and Spanish prose
Caliskan-Islam 2015 [33]	Lexical and syntactic features	Random Forest	Source code

- The approaches discussed thus far have mostly been applied to problems in which a large amount of text is available (e.g., novels, essays, newspaper articles, etc.).
- SVM apply directly, better performance can be achieved with features and classification approaches custom-tailored for attribution problems with small samples of texts.
- More specific to Internet messaging, they also evaluated structural attributes of the messages including the presence of a greeting, farewell, and signature in the text.
- Combined with SVM for classification, these features were shown to be reasonably effective for attribution problems consisting of a small number of authors.

- The idea behind author unmasking is that the differences between two texts from the same author will be reflected in a relatively small number of features.
  - It was shown that there is significant aliasing between different-author and same-author performance curves when considering samples of 5,000 words or less.
  - The sequence-based approaches of Character-level Markov Chains and Character-level Sequence Kernels are suggested as alternatives.
  - When features from partial parsing are combined with SVM, high accuracy can be achieved (over 90%) for samples as small as 200 words

❑ Programming

- programming language authorship attribution, the state of the art for de-anonymizing programmers shows significant promise for the analysis of short samples of text.
- The best reported methodology makes use of features derived from abstract syntax trees.
- Caliskan applied this methodology to code stylometry
- Information gain was applied to select only the more informative features, making the approach more accurate and the computation more tractable.
- When validated on code solutions that were, on average, 70 lines of code long, accuracies for distinguishing between sets of 1,600 and 250 programmers reached 94% and 98%.

II.C :Authorship Attribution for Small Sample of Text

- The authorial style of malicious source code often percolates through other media where short-form writing is prevalent.
- Afroz conducted a large-scale study of posts on undergroundforums relatedto passwordcracking,spam, credit card fraud, software exploits, and malicious search engine optimization.
- Afroz et al. suggest several that are amenable to SVM classification character-level unigrams and tri-grams, word-level bi-grams, numbers used in place of letters, capitalization, parts of speech, and the presence of foreign words.

II.C :Authorship Attribution for Small Sample of Text

TABLE IX
WORKS IN SHORT TEXT AUTHORSHIP ATTRIBUTION (SORTED BY TYPE OF CORPUS)

Source	Features Used	Classifier	Corpus
Sanderson and Guenter 2006 [158]	Character and word sequences	Character-level Sequence Kernel, Markov chains and SVM	English short text samples
Hirst and Feiguina 2007 [74]	Syntactic labels	SVM	English short text samples
Koppel et al. 2007 [111]	Function words	SVM	English essays
Forstall et al. 2011 [56]	Character-level n-grams	SVM	Latin poetry
Anderson 2001 [6]	Capitalization, white space, and punctuation	SVM	English e-mail
de Vel et al. 2001 [46]	Capitalization, white space, and punctuation	SVM	English e-mail
Koppel and Schler 2003 [106]	Function words, part-of-speech tags, idiosyncratic usage	Decision Trees and SVM	English e-mail
Layton et al. 2012 [115]	Character-level n-grams	SCAP	English electronic communication
Brocardo et al. 2013 [27]	Character-level n-grams	*Ad hoc* similarity measure	English e-mail
Koppel et al. 2011 [109]	Character-level n-grams	Cosine similarity	English web posts
Koppel and Winter 2014 [113]	Character- and word-level n-grams	SVM	English web posts
Qian et al. 2014 [152]	Word and character-level n-grams and syntactic features	SVM	English web posts
Afroz et al. 2015 [3]	Lexical, syntactic and domain-specific features	SVM	Russian, English and German web posts
Frantzeskou et al. 2006 [60]	Byte-level n-grams	SCAP	Source code
Frantzeskou et al. 2007 [59]	Byte-level n-grams	SCAP	Source code
Hayes 2008 [72]	Lexical features	Multivariate analysis and Linear Discrim. Analysis	Source code
Burrows and Tahaghoghi 2007 [32]	Token-level n-grams	Statistical hypothesis test	Source code
Caliskan-Islam et al. 2015 [34], [35]	Lexical and syntactic features	Random Forest	Source and compiled code

II.D : Authorship Attribution Specifically for Social Media

- A growing body of work has attempted to mine and analyze actual online postings.
- A Karhunen-Löeve transform-based technique dubbed “WritePrints” was applied to this data, showing accuracy as high as 94% when differentiating 100 distinct authors
- To date, only a handful of researchers have tackled the authorship attribution problem for tweets collected in the wild using the techniques described above.
- Methods relying on SVM for classification outperform other approaches to authorship attribution on tweets, namely Naïve-Bayes, Source-Code Authorship Profiling, and other simple similarity measures.
- Almost all of these approaches used the same set of features, character-and word-leveln-grams.

- A further complication is the need for automatic language understanding for Eastern and Near Eastern language posts, where non-Latin character sets are used, and individual characters may express entire words.
- While language-dependent approaches like partial parsing fail without substantial retuning in such cases, language-independent character-level n-gram and TF-IFD-based approaches work just fine for non-Latin characters with no adjustment.

- One might also ask if a writing style evolves over time in a way that is unique to a specific social media platform like Twitter.
  - Azarbonyad et al. have studied this question, and have isolated distinct vocabulary changes for the sam e authors of tweets over a period of months.
  - The cause might be as simple as a change in the circumstances of an author's life, or as nuanced as the absorption of stylistic traits after reading the tweets of others.
  - It is possible to design a time aware attribution algorithm that constructs a language model for distinct periods from an author'scollected messages, which can be achieved by calculating decay factors that are applied as weights to the periods.
  - Using character-level n-grams as a feature basis, Azarbonyad et al. showed that a time-aware SCAP approach is far more effective than a baseline without any temporal weighting.

- Going beyond lexical- and sound-oriented features, semantic analysis can also be applied to attribution tasks.
- DADT(Disjoint Author-Document Topic Model)
  - author topics are disjoint from document topics, different priors are placed on the word distributions for author and document topics, and a ratio between document words and author words is learned.
  - The feasibility of this approach has been demonstrated on emails and blog posts.
  - However, it is not always possible to perform meaningful semantic analysis on sample sizes as small as tweets with any of today's topic modeling algorithms.
- By aggregating tweets into per-user profiles for training and testing, conventional topic modeling algorithms can be applied with little trouble.
  - This strategy is not feasible if we are considering just a single testing tweet in an actual investigation.

II.E : The Threat of Counter-Attribution

- Naturally, in authorship attribution, there exists some element of the "offense and defense" dynamic present in the broader world of computer security.
- Counter-attribution techniques
  - where there is an intentional act of changing one's writing style, have emerged to thwart authorship attribution systems.
  - counter-attribution can be misused by malicious actors attempting to evade identification by various authorities.
- Kacmarcik and Gamon describe shallow anonymization, whereby 14 changes per 1,000 words disrupts an SVM classifier trained with function words, whereby increasing numbers of feature modifications defeat an approach relying on the rate of degradation of the accuracy of learned models.

II.E : The Threat of Counter-Attribution

- Juola and Vescovi studied the impact of the Brennan-Greenstadt corpus , which was cleverly crafted to deliberately mask style, on the Java Graphical Authorship Attribution Program.
- The Brennan-Greenstadt corpus makes use of the following strategies for counter-attribution:
  - obfuscation (i.e., identity hiding)
  - imitation (i.e., the deliberate misappropriation of another author's style)
  - translation (i.e., the use of machine translation to alter an author's signal)
- The findings of Juola and Vescovi indicate that all common feature sets are impacted by counter-attribution.

II.E : The Threat of Counter-Attribution

- Fundamentally, if an author's signal is scrubbed from a text, we would have to turn to other evidence associated with the case to make a confident attribution decision.
- It remains unknown whether or not a perfect counter-attribution solution can be developed, given the vast linguistic feature space of any underlying text.

III. Walk-through of authorship attribution techniques for social media forensics

A. General framework for authorship attribution
B. Source data from twitter
C. Bag-of-words model
D. Classification strategies

- This section will walk through the process of identifying the author of a given set of tweets from twitter.

- Basic strategy : Relies on a set of features capturing patterns extracted from the original texts in **a bag of words model** dynamically created for a set of user.

- **Method of creating a bag of model**
  - **one model for a set of authors (저자 집단당 1개의 모델)**
    - Overlook some discriminative features of particular authors ➔ may not be strong enough to appear globally
  - **one model per author (저자 당 1개의 모델)**
    - Allow for a more fine-grained stylometric evaluation ➔ as the number of authors increases, creating dynamic model for each investigated author is much more consuming.

In this paper, for computational efficiency, consider the case of a model per set of authors.

- Various classification method are applicable to this problem.
  - Most the Support Vector Machine is used to address high-dimensional feature representations.

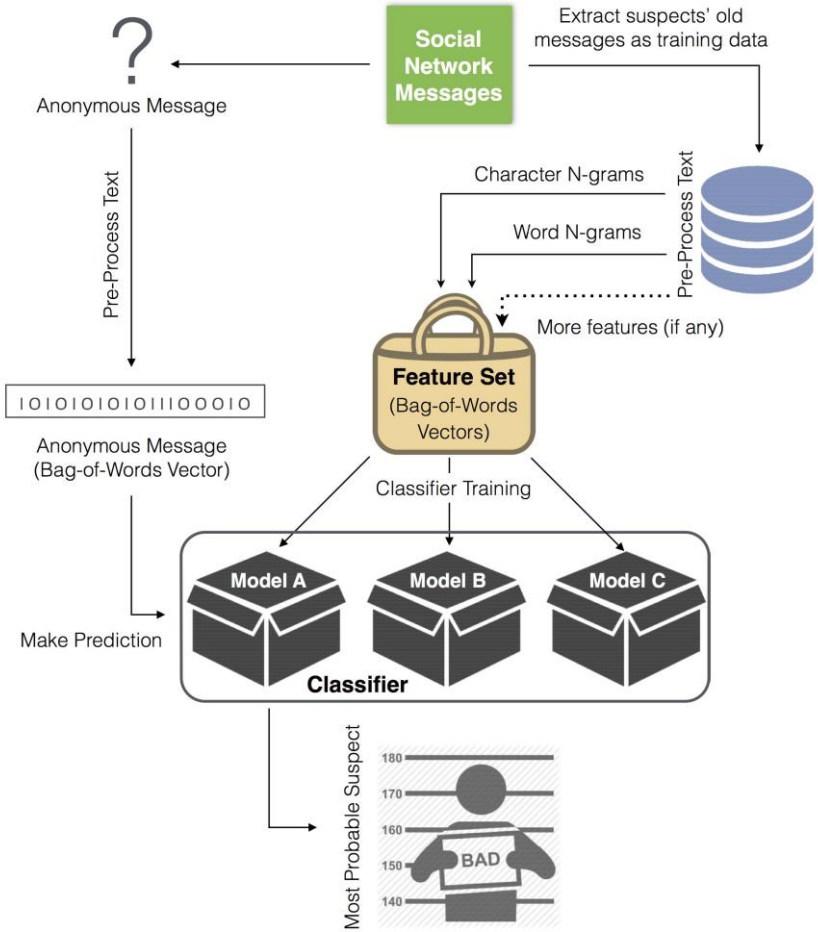
- **Interest of classifier**
  - better handle large-scale with respect to accuracy, as well as speed of processing
  - For open set attribution problems, classifiers specifically designed to mitigate the risk of the unknown

❑ General framework

- Recommend an approach that scales well with a large number of suspects

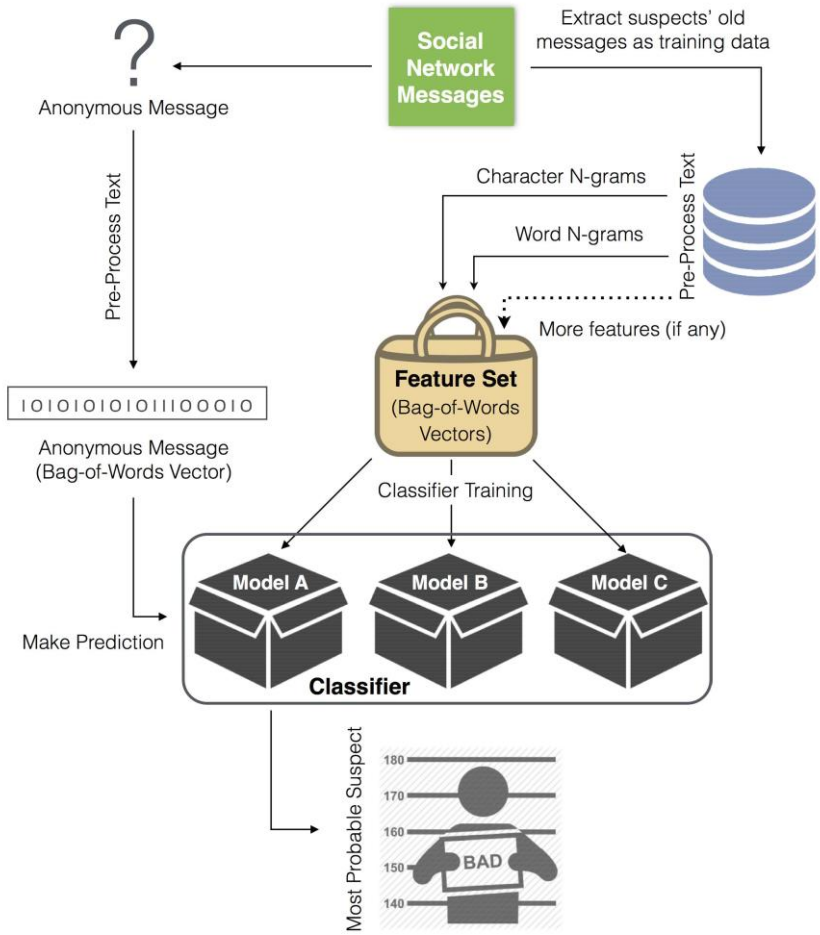
❑ Typical scenario

- A tweet describes illicit activity, sent anonymously from within a small company
- If there is no physical evidence linking one of those employees to the message, all employees become suspects.
- The use of a machine learning approach becomes paramount



❑ Step of the authorship attribution system in this paper

- Step1. Messages are harvested from the social media accounts of known suspects
- Step2. After enough examples have been collected, the raw data is pre-processed
  - remove very sparse features, very shot message, and non-English messages ➔ **enforces consistency**
- Step3. Evaluate character-level N-grams, word-level N-grams, part-of-speech N-grams, and diverse lexical and syntactic statistics as features
- Step4. Collect the above features to features set based on Bag-of-Words
- Step5. Train the classifier use various classification method
  - The power mean support vector machine(PMSVM)
  - W-SVM
  - Random forest
  - SCAP (source code authorship profiling)
  - compression based attribution
- Step6. Test the anonymous message
- Step7. Using the pretrained classifier, make prediction of its authorship



III.B: Source data from twitter

❑ Data extraction

- The ultimate forensics goal : authorship attribution ➔ all the retweet should be removed
  (using the Twitter API, we remove the retweet flag and meta tag RT.)
- This paper focus is on English-language tweets, thus non-English tweets can be removed using the python library 'guess-language'
- Using the spell-checking python library 'pyenchant' to build an accurate prediction of the language of a text consisting of three or more words
  ➔ Recommend all messages that contain only one or two words.
  Because shot messages don't provide meaningful information about the author and end up introducing noise into the classification task

❑ Text pre-processing

- There is no need for strong pre-processing in authorship attribution
- Author's writing style is an integral part
  ➔ it does not make sense to stem words or correct the grammar of messages under consideration.
- Therefor instead of more elaborate tokenization our pre-processing will focus on normalizing very sparse characteristics.
  - numbers, dates, times and URLs
  - Hashtag, user reference
    ➔ it makes supervised learning method unreliable, because a user might make references to the same person across her messages, creating a strong bias towards that particular feature in training and any message with a reference to that same person would subsequently be misclassified as being from that user.

❑ Text pre-processing example

Before pre-processing

Tweet 1 : "Do not forget my bday is on 03/27 #iwantgifts"

Tweet 2 : "@maria I will be sleeping @00:00AM"

Tweet 3 : "Check out this amazing website: http://www.ieee.org"

After pre-processing

Tweet 1 : "Do not forget my bday is on DAT TAG"

Tweet 2 : "REF I will be sleeping @TIM"

Tweet 3 : "Check out this amazing website: URL"

❑ Bag-of-Words model

- The bag-of-words is a classic model in natural language processing.
- It is an orderless document representation of feature frequencies from a dictionary.

❑ Example

Text message

Text 1 : "To be, or not to be, that is the question."
Text 2 : " To die, to sleep. To sleep, perchance to dream."

Mapping dictionary

#	1	2	3	4	5	6
Word	to	be	or	not	that	is
#	7	8	9	10	11	12
Word	the	question	die	sleep	perchance	dream

feature

Feature vector

F.V 1 : [1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0]
F.V 2 : [1, 0, 0, 0, 0, 0, 0, 0, 1, 1 ,1 ,1]

❑ Character-Level n-Grams

- Often used for authorship attribution in texts from social media
  - They cans capture unusual features, such as emoticons and special use of punctuation.
  - They help mitigate the effect of small typos that authors do no repeat very often, which are not style marker.
    - **For example**

misspeling : miss, issp, sspe, spel, peli, elin, ling	misspelling : miss, issp, sspe, spel, pell, elli, llin, ling

- This paper focus on character-level 4-grams, with whitespace and meta tags included in the n-gram
  - With respect to Twitter, whitespace is appended at the beginning and at the end of each tweet
- Discard any character-level 4-gram which does not appear at least twice for the same author in the straining set, thus eliminating hapax legomena
  - Improves efficiency by removing noisy features that are unlikely to appear again in the future for the same user
  - Hapax legomena : features that occur exactly once in a particular setting
- The features are case-sensitive
  - The author's preference for capitalization of letters is also one of the traits that can be used for attribution
  - Many user of social media have a preference for capitalizing some words to emphasize them

❑ Character-Level n-Grams example

Text message

"2B!!!! or n2B!!!!!!! ;)"

Char. Level 4-grams

1	2	3	4	5	6	7	8	9	10	11	12	13	14
_2B!	2B!!	B!!!	!!!!	!!!_	!!_o	!_or	_or_	or_n	r_n2	_n2B	n2B!	2B!!	B!!!
15	**16**	**17**	**18**	**19**	**20**	**21**	**22**						
!!!!	!!!!	!!!!	!!!!	!!!_	!!_;	!_;)	_;);						

❑ Word-Level n-Grams

- Word-level n-grams let us capture more semantically meaningful information from a text in the form of short phrases.
  - When considering messages from social media, the constraints placed on message length force users to be more judicious in their writing.
  - Thus, it is reasonable to assume that authors will economize, and only repeat very short phrases.

- A good rule of thumb
  - To use word-level n-grams where $n \in \{1, \ldots, 5\}$

- Length of feature vectors
  - The choice of N
  - The number of known authors
  - the number of texts considered during training
  - For instance, when n = 4 the feature dimensionality varies from 20,000-dimensional vectors (50 users and 50 training tweets per user) to around 500,000-dimensional vectors (500 users and 500 tweets per user).

❑ Word-Level n-Grams (cont.)

- Although some researchers have argued for n < 4 as a default setting, for messages from social media, we need a larger n to capture the idiosyncratic language usage of the Internet, which includes emoticons, onomatopoeia (a word that resembles its associated sound, e.g. cuckoo), abbreviations, and other unconventional usage.

❑ Word-Level n-Grams example

Text message

"To be, or not to be, that is the questions."

Word level-1-gram(Unigram)

1	2	3	4	5
To	be	or	not	to
6	**7**	**8**	**9**	**10**
be	that	is	the	question

Word level-2-gram(Bigram)

1	2	3	4	5	
BEGIN To	To be	be or	or not	not to	
6	**7**	**8**	**9**	**10**	**11**
to be	be that	that is	is the	the question	question END

Word level-3-gram(Trigram)

1	2	3	4	5	6	7	8	9	10
BEGIN To be	To be or	be or not	or not to	not to be	to be that	be that is	that is the	is the question	the question END

❑ Part-of-Speech (POS) n-Grams

- The simplest stylistic features related to syntactic structure of texts are part-of speech(POS) n-grams.
- POS tagging is a process that can be performed easily and with relatively high accuracy.
  - Given a POS tagger that has been trained with texts possessing similar properties to the ones under investigation, it can achieve near-optimal accuracy at the token level
  - Noise in stylistic measures can significantly affect the performance of the attribution model
- POS tag
  - One way to mitigate this concern because they reduce the feature-space to a limited number of very general elements.
  - In this study, we use a POS tagger specifically designed for tweets in English. A POS tag-set of 25 tags, covering twitter-specific features like hashtags, at-mentions, retweets, URLs and emoticons was used.

❑ POS-Level n-Grams example

Text message

"Thx 4 having me. Great time w/ you all"

PoS tags

1	2	3	4	5
N	P	V	O	W
6	**7**	**8**	**9**	**10**
A	N	P	O	D

TABLE V
PART-OF-SPEECH TAGS CONSIDERED IN THIS WORK (ADAPTED FROM [65])

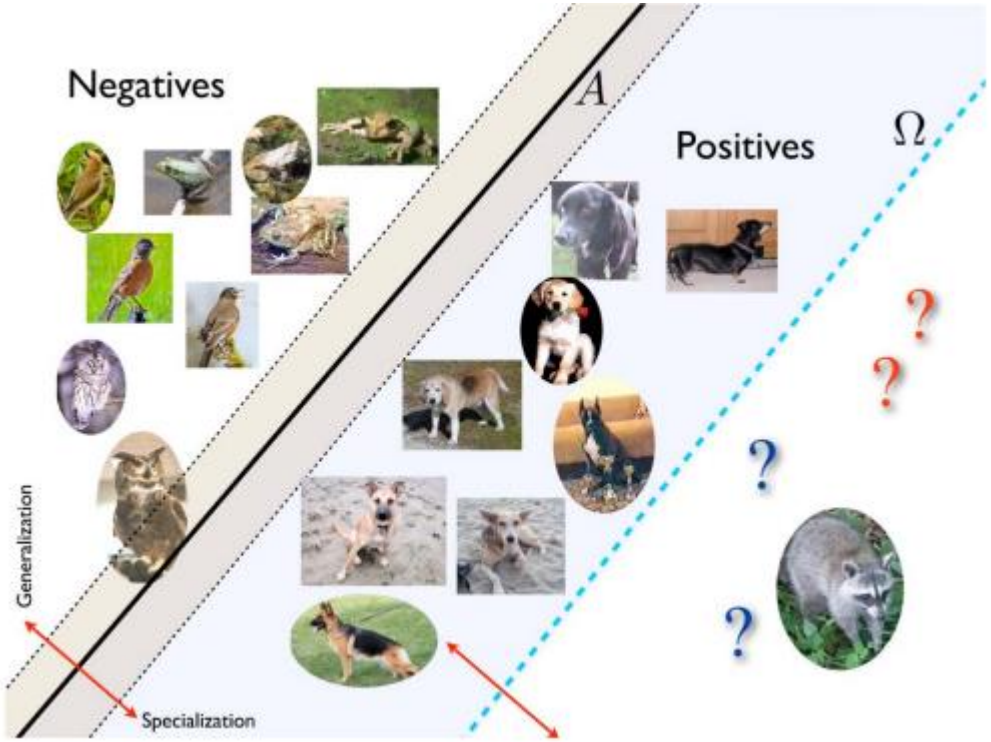
Tag	Meaning
A	adjective (J*)
B	proper noun (NNP, NNPS)
C	interjection (UH)
D	determiner (WDT, DT, WP$, PRP$)
E	emoticon
F	coordinating conjunction (CC)
G	other abbreviations, foreign words, possessive endings, symbols, garbage (FW, POS, SYM, LS)
H	hashtag (indicates topic/category for tweet)
I	at-mention (indicates another user as a recipient of a tweet)
J	discourse marker, indications of continuation of a message across multiple tweets
K	numeral
L	nominal + verbal
M	proper noun + verbal
N	common noun (NN, NNS)
O	pronoun (personal/WH; not possessive; PRP, WP)
P	pre- or postposition, or subordinating conjunction (IN, TO)
R	adverb (R*, WRB)
S	nominal + possessive
T	verb particle (RP)
U	URL or email address
V	verb incl. copula, auxiliaries (V*,MD)
W	punctuation
X	existential there, predeterminers (EX, PDT)
Y	X + verbal
Z	proper noun + possessive

❑ A more diverse feature set for open set attribution

- Open set recognition problems are among the hardest in machine learning.
- A problem such as open set authorship attribution is difficult because there are often small interclass distances in the feature space – in some cases, an author's style is very similar to that of other authors, as a function of education, influence or culture.

❑ Authorship verification

- Given two tweets, a prediction is made as to whether or not they are from the same author.
  - effective for closed set attribution does not yield enough information diversity to make accurate predictions.
- Thus we must turn to **feature-level fusion** over more diverse feature types to capture additional facets of style at the character and word levels
  - feature-level fusion : the process of combining different individual features that have been extracted from the input text samples into one feature vector before training or classification



❑ Power Mean SVM (PMSVM)

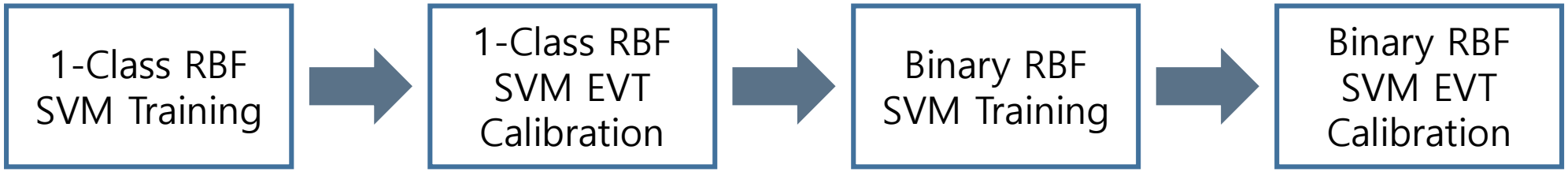
- PMSVM was originally proposed for large-scale image classification.
- These kernels directly apply to applications such as image and text classification, where the data is well represented by histograms or bag-of word models.
- Generalize many kernels in the additive kernel family
- This kernel family is not very sensitive to parametrization, avoiding overfitting to the training data.
- Aggregate the advantages of linear SVM and non-linear additive kernel SVM
- It performs faster than other additive kernels because, rather than approximating the kernel function and the feature mapping, it approximates the gradient function using polynomial regression.

$$M_p(x_1, \ldots, x_n) = \left(\frac{\sum_{i=1}^{n} x_i^p}{n} \right)^{\frac{1}{p}} \qquad M_p(\vec{x}, \vec{y}) = \sum_{i=1}^{d} M_p(x_i, y_i).$$

- This formulation would lead to higher training times, but the PMSVM algorithm uses the coordinate descent method with a gradient approximation to solve the dual SVM problem.
- Training is also faster and the approximation avoids overfitting to the training data.

W-SVM for open set attribution

- A Weibull-based formulation that combines a 1-Class SVM with a binary SVM
- Reasons to help for open set attribution
  - A binary model gives us an explicit class for rejection in the authorship verification scenario
  - When Weibull modeling is coupled with a 1-Class SVM with a radial basis function(RBF) kernel, it can be proved that the probability of class membership decreases in value as points move from known training data toward open space
  - The Weibull distribution provides better probabilistic modeling at the decision boundary for a binary SVM



❑ Random forests

- a method which comprises a collection of classification or regression trees, each constructed from a random resampling of the original training set.

❑ Source Code Author Profile (SCAP)

- Profile-based methods first concatenate all available training tweets per author and then extract a single representation from them attempting to collectively describe the author's profile.
- Since this training phase is very simple, an inherent advantage of profile-based approaches is that they can easily be scaled to additional candidate authors.

❑ Step of SCAP

- SCAP builds a profile for each author that is based on the k most frequent character-level n-grams in the texts of that author.
- Each evaluation text is also represented using the list of its k most frequent character-level n-grams, with attribution decisions made based on which author it shares the most n-grams with.
- This article, n=4 used

❑ Compression-based attribution

- The main idea is that a text of unknown authorship is more likely to be effectively compressed with other texts of its true author rather than with texts of other authors.
- Such an approach can easily be implemented using off-the-shelf text compression algorithms like rar, bzip2, and gzip.
- Compression-based methods do not extract a concrete representation of texts with clearly defined features.
- They are usually based on character sequences repeatedly used within texts, and can identify common patterns between the unknown texts and the candidate authors.

❑ Prediction by Partial Matching (PPM)

- to compress the concatenation of all available training texts per author.
- Then, for a given document of unknown authorship D, it calculates the document cross-entropy that corresponds to the average number of bits per symbol to encode the document using the author's model
- The candidate author that minimizes document cross-entropy is the most likely author of D.

❑ Compression-based attribution

- The main idea is that a text of unknown authorship is more likely to be effectively compressed with other texts of its true author rather than with texts of other authors.
- Such an approach can easily be implemented using off-the-shelf text compression algorithms like rar, bzip2, and gzip.
- Compression-based methods do not extract a concrete representation of texts with clearly defined features.
- They are usually based on character sequences repeatedly used within texts, and can identify common patterns between the unknown texts and the candidate authors.

❑ Prediction by Partial Matching (PPM)

- to compress the concatenation of all available training texts per author.
- Then, for a given document of unknown authorship D, it calculates the document cross-entropy that corresponds to the average number of bits per symbol to encode the document using the author's model
- The candidate author that minimizes document cross-entropy is the most likely author of D.

IV. Experimental result

A. Data set and pre-processing
B. Comparison of different feature types
C. Comparison of different classifiers
D. Efficiency and search space reduction
E. Feature importance
F. Open set attribution

❑ Viable validation regime

- Schwartz et al. [166] introduced a viable validation regime for authorship attribution methods targeted at social media.
- **Two important aspects of the problem**
  - The impact of varying training set sizes
  - The impact of varying number of authors

1) A comparison of the performance of various feature types using a fixed pool of 50 Twitter users using PMSVM and Random Forests classifiers

2) A comparison of the performance of various classifiers by varying the number of Twitter users and feature types

3) An assessment of algorithm efficiency and search-space reduction

4) An analysis on feature importance given a fusion method using different groups of features

5) A comparison of different methodologies for open set attribution.

IV.A: Data set and pre-processing

❑ Public data set

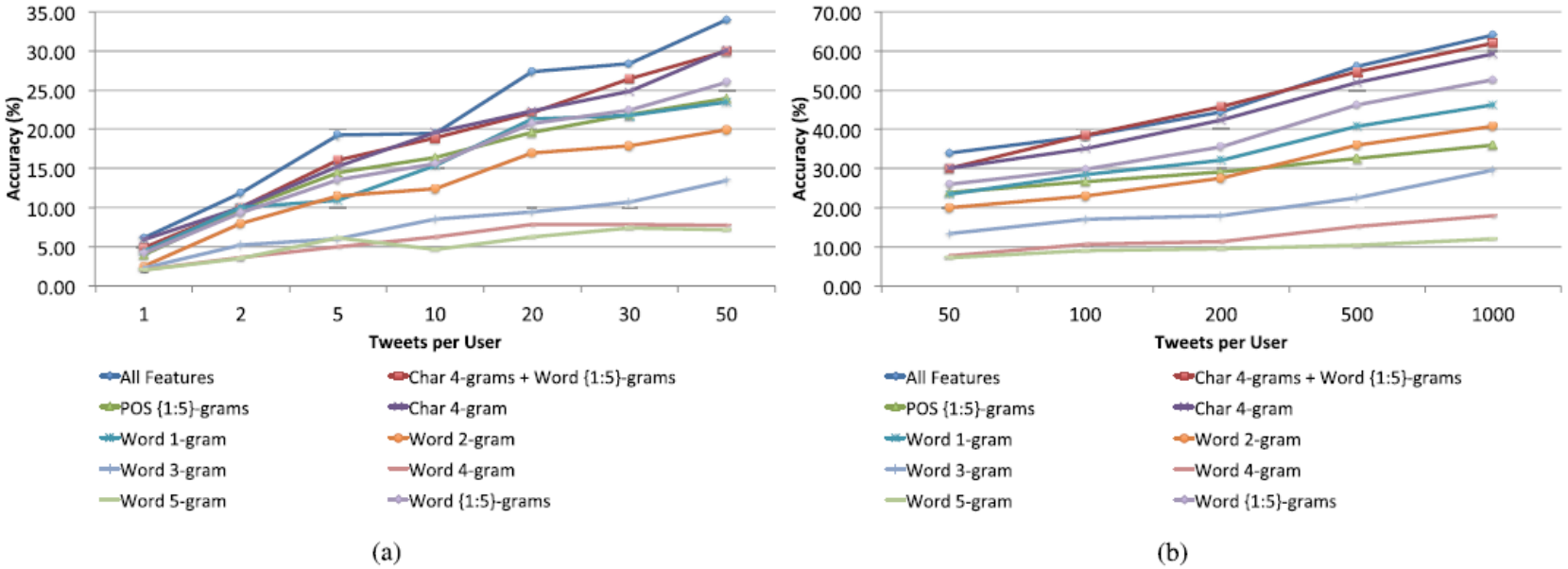
- Not exist for authorship attribution applied to social media forensics
- Also, the restrictive terms of use put in place by the major social networks prohibit the dissemination of such data sets.

❑ Data set

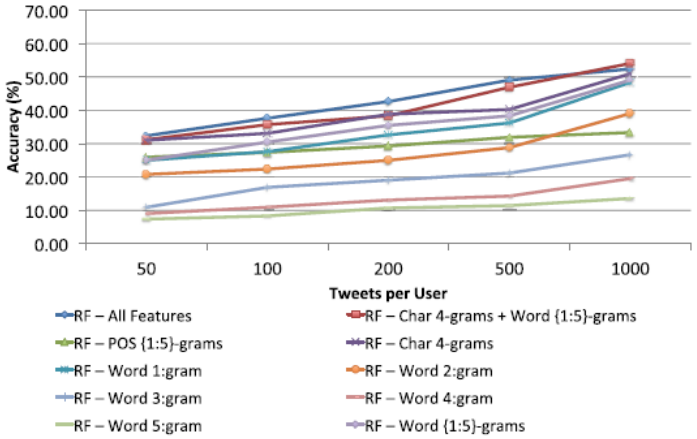
- Authors created our own large-scale data set that was designed with algorithm scalability evaluations in mind.
- Collected ten million tweets from 10,000 authors over the course of six months in 2014.
  - Each tweet is at most 140-character long and includes hashtags, user references and links.
- Pre-processing of each tweet includes removing all non-English tweets, tweets with less than four words, and tweets marked as retweets or any tweet containing the meta tag RT.
- The data set was partitioned into training and test sets via k-fold cross validation.
  - IV-B, IV-E : 10 fold cross validation
  - IV-F : 5 fold cross validation

IV.B: Comparison of different feature types

❑ PMSVM



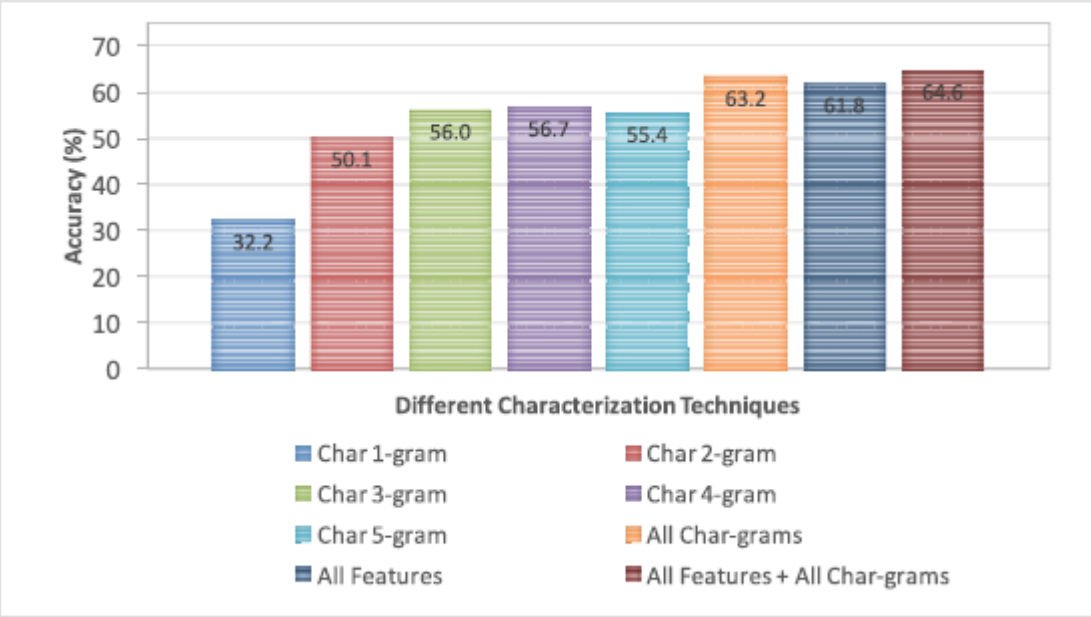
❑ Random forests



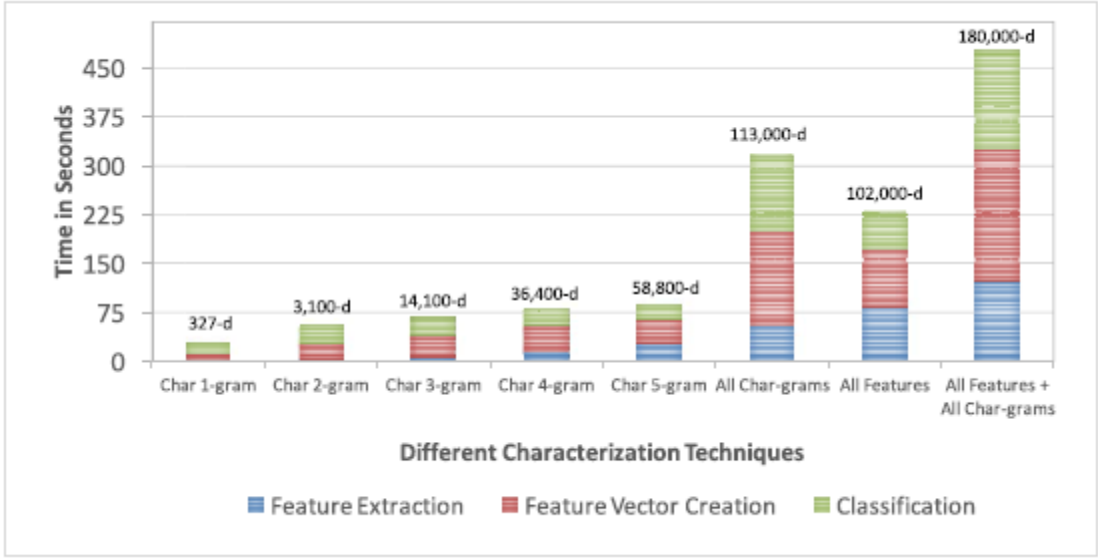
- The classification using only a few micro messages is still an open problem with the performance steadily improving as more messages are present in the training pool.

- The figure also shows that the unigrams are relevant features.

IV.B: Comparison of different feature types

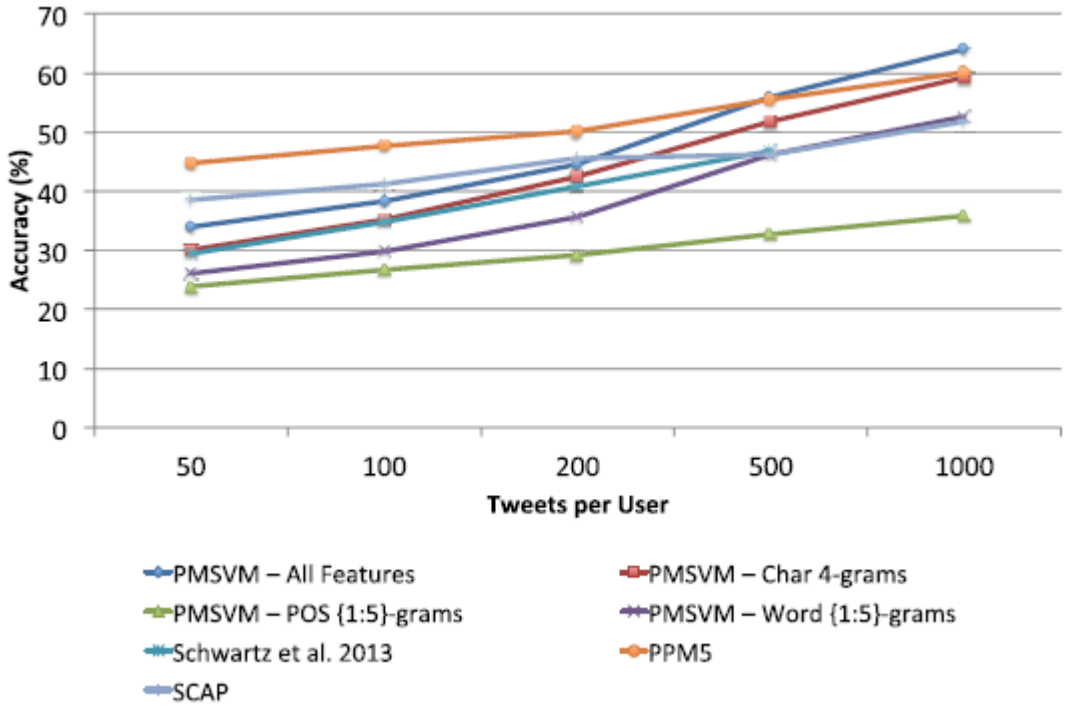


Effects of choosing different character n-grams and their combination.

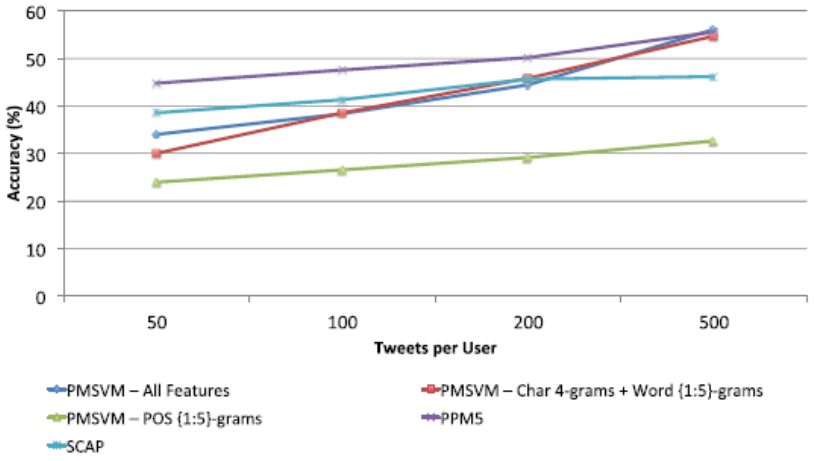


Training time breakdown into three major tasks: feature extraction, vector creation and classification learning. On top of each stacked bar, we show the final average feature vector length for each method.

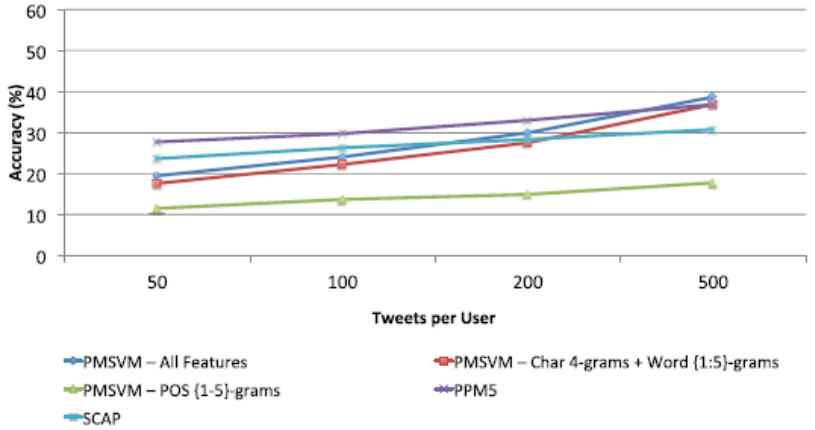
IV.C: Comparison of different classifiers



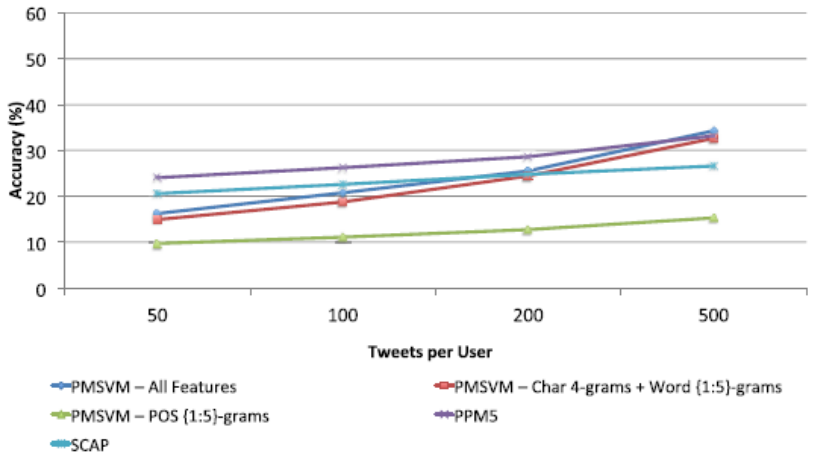
IV.D: Efficiency and search-space reduction



(a) author : 50

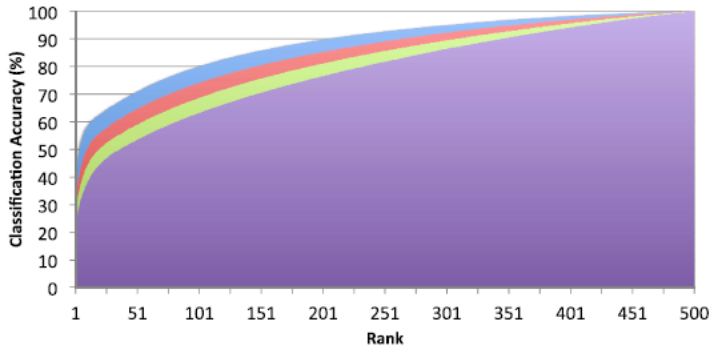


(b) author : 500



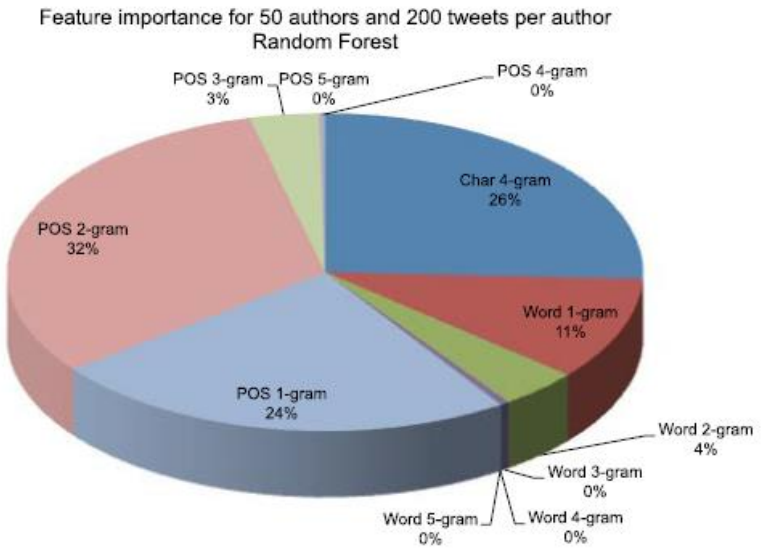
(c) author : 1000

- This shows that investigators could keep adding training examples if they exist in an attempt to improve accuracy — something that is not possible with all classifiers.



Cumulative Match Curves for PMSVM for 500 Twitter users with a varying number of tweets per user

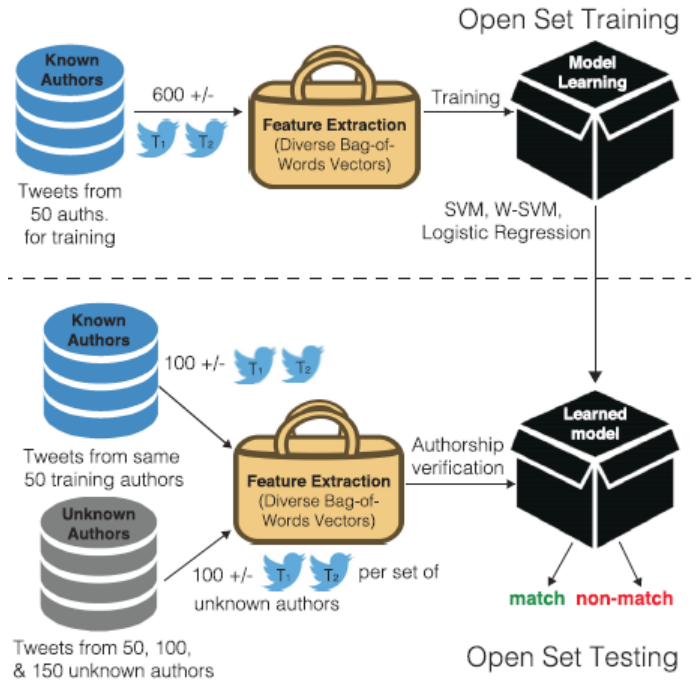
IV.E : Feature Importance

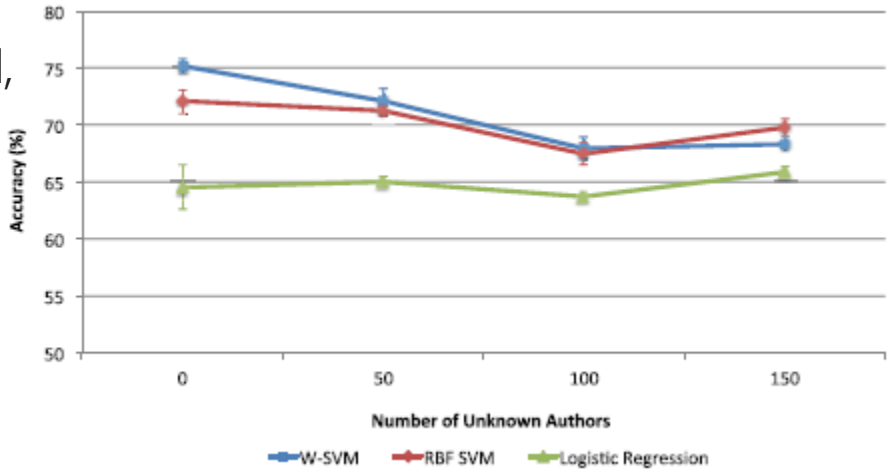


Feature importance determined by a Random Forest classifier for a fused vector consisting of different word-level n-grams, character-level 4-grams, and different POS n-grams for 50 authors and 200 training tweets per author. Note that the entries marked as 0% are actually very close to zero and not literally zero.

An overview of open set authorship verification experiment

- 50 known authors were randomly chosen from the corpus and fixed across folds for this experiment.
- For each fold, positive training samples for these authors were generated by randomly sampling 600 matching tweet pairs (*i.e.*, two different tweets from the same author), and negative samples generated by sampling 600 non-matching tweet pairs (*i.e.*, two different tweets from two different authors).
- This training data was used to create three separate verification models that can tell if any two tweets came from the same author.
- For the models, the open set-specific W-SVM classifier was considered, along with RBF SVM and Logistic Regression for comparison.
- Parameters for each classifier were tuned via cross-validation during training.





V. Conclusion and future directions

A. Real-world use
B. Social network scale data
C. Dense features
D. Representativeness
E. Open set recognition
F. Decision level fusion

V : Conclusion

- The enormous popularity of social media means that it is now a conduit for both legitimate and illegitimate messages targeted at the broadest possible audience.
- new forensic challenges have appeared related to this form of new media.
  - A primary problem in this area has been authorship attribution for short messages.
- This study showed that for popular services like Twitter, we face the dilemma of simultaneously having an enormous overall corpus, yet scarcity of information for individual users.
  - we should consider strategies that are a bit different than traditional authorship attribution algorithms for long form writing.
  - One way to address this problem is to compute very low-level lexical statistics.

V : Conclusion

- There are several complementary paths one can follow when solving the authorship attribution problem for social media forensics.
  - The PMSVM algorithm used in conjunction with diverse and complementary features is a good advance over the state-of-the-art methods.
  - The cumulative matching analysis showed that current attribution methods can greatly reduce the number of users to be analyzed in a real situation.

❑ Real-world use

- The methods we have discussed can be used directly by investigators, especially for suspect search-space reduction or helping to assemble more conventional evidence that is admissible in court.
- For further clues, researchers should also look closely at the graphical structure of a social network.
- Such an analysis might surface additional suspects that are implicated via their association to an identified author.

❑ Social network-scale data

- The experiments in this article indicate that by using more training data, better results can be achieved.
  - Methods like PMSVM, which are custom-tailored to high-dimensional spaces, represent a large improvement over prior method using traditional learning.
- Future work should also look at meta-data, images, and videos posted online to further hone the accuracy of attribution by pursuing a hybrid approach that extends beyond just stylometry.

❑ Dense Features

- The methodology we introduced consisted of dynamic features extracted from raw text, and and messages like tweets contain few words.
- Standard techniques such as PCA and random selection of features to reduce the dimensionality of the feature vectors.
  - Random selection always performed worse, no matter the size of the extracted features.
  - PCA failed to converge in some cases due to the overwhelming number of examples and dimensions and the limits of the available hardware.
- The reason for the drop in accuracy and speed
  - information may be discarded when reducing dimensionality
  - interplay of the resulting dense representation and the classification algorithm processing it.

V.D : Representativeness

- Unlike other text classification tasks, in authorship attribution it is not always possible to assume that training and evaluation texts will share the same properties.
  - For instance, in a forensic examination of suspicious tweets, for some of the suspects we may not find authentic tweets by them.
- However, we might be able to find other types of text like email messages, blog posts, etc. Certainly, the available texts by each suspect may be on a completely different topic in comparison to the texts under investigation.
  - In all these cases, the training dataset is not representative enough with respect to the documents under investigation.
- Thus, unlike other text classification tasks, we need authorship attribution models that remain useful in cross-genre and cross-topic conditions.
- But the lack of large-scale data covering multiple and heterogeneous genres and topics by the same authors limits the generality of the conclusions.
- Another problem that affects the representativeness of the training data is the distribution of that data over the candidate authors.

V.E : Open set recognition

- A traditional multi-class classifier will always return an answer pointing to a known suspect, which will not be correct in many instances.

- This suggests that the effort should be placed on reducing and prioritizing the known suspects rather than always pointing to a single culprit.

- In this review, authors introduced a basic methodology for open set attribution incorporating an algorithm that is known to minimize the risk of the unknown, and presented an experiment that highlighted the difficulty of the problem.

V.F : Decision level fusion

- In this paper, authors investigated combining different character- and word-level n-gram feature set along with POS-tag n- grams, which turned out to be effective

- However according to analysis, not all features are equally important, which shows that investigating how each group of features affects the performance is an important research direction.
  - For instance, combining the output of the PMSVM, Random forest, SCAP and PPM models to ideally arrive at a more accurate result, compared to examining each model in isolation

THANK YOU