· Universitá degli Studi di Verona ·

FACOLTÁ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea Magistrale in INGEGNERIA E SCIENZE INFORMATICHE
Curriculum VISUAL COMPUTING

STATISTICAL ANALYSIS OF SKYPE CONVERSATIONS: RECOGNIZING INDIVIDUALS BY THEIR CHATTING STYLE

Tesi di Laurea Magistrale

Relatore	Controrelatore	Presentata
Dott.	Dott.	da:
MARCO CRISTANI	MATTEO CRISTANI	CRISTINA SEGALIN

Sessione Laurea Estiva
Anno Accandemico 2011/2012
Consultazione consentita

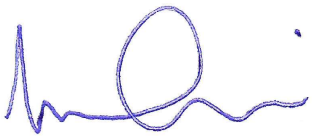
· Universitá degli Studi di Verona ·

FACOLTÁ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea Magistrale in INGEGNERIA E SCIENZE INFORMATICHE
Curriculum VISUAL COMPUTING

STATISTICAL ANALYSIS OF SKYPE CONVERSATIONS: RECOGNIZING INDIVIDUALS BY THEIR CHATTING STYLE

Tesi di Laurea Magistrale

Relatore	Controrelatore	Presentata
Dott.	Dott.	da:
MARCO CRISTANI	MATTEO CRISTANI	CRISTINA SEGALIN



Sessione Laurea Estiva
Anno Accandemico 2011/2012
Consultazione consentita

Master Degree thesis completed between May and June 2012 in the VIPS Lab. of the Computer Science Department.

Website:
http://criistis.wordpress.com

E-mail:
criistina.segalin@gmail.com

Dedicated to all researchers, hoping for a better University in Italy.

"The nature of an innovation is that it will arise at a fringe where it can afford to become prevalent enough to establish its usefulness without being overwhelmed by the inertia of the orthodox system."

— K.Kellys

Abstract

Authorship attribution (AA) aims at recognizing automatically the author of a text sample. Traditionally applied to literary texts, AA faces now the new challenge of recognizing the identity of people involved in chat conversations.
These share many aspects with spoken conversations, but AA approaches did not take it into account so far. Hence, we try to fill the gap proposing two novelties that improve the effectiveness of traditional AA approaches for this type of data: the first is to adopt features inspired by Conversation Analysis (in particular for turn-taking), the second is to extract the features from individual turns rather than from entire conversations. In conversations, turns are intervals of time during which only one person talks. In chat interactions, a turn is a block of text written by one participant during an interval of time in which none of the other participants writes anything.

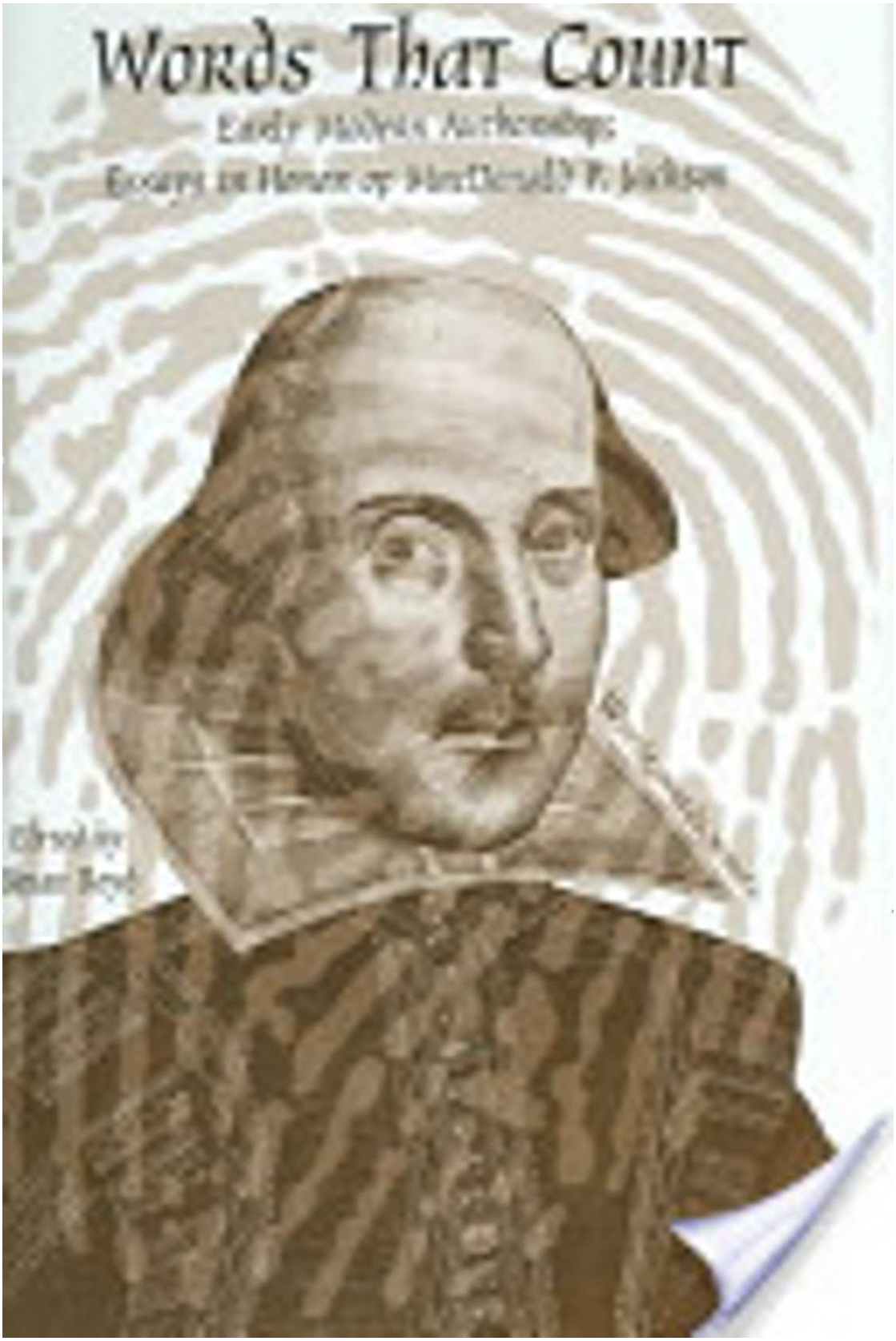
To face this challenge we try to use both verbal and non-verbal cues, extracted from a corpus of dyadic chat conversations (77 individuals in total). The verbal cues can be called stylometric features or writeprints, as they are similar to fingerprints, but composed of multiple features, such as vocabulary richness, length of sentence, function words, layout of paragraphs. Considering the turn-taking of a chat as if it was a spoken conversation, we have to take into account also non-verbal cues as turn duration, response time, writing speed that we called conversational cues.

In the modeling, we preferred to keep the most informative features that occurred most of the times in the raw feature set extracted form chat conversations. The feature extraction does not consider the content of the message for privacy and ethical issues limits. Considering 77 individuals, the probability of finding the right match among the first N subjects is 89.5. We also try to establish if, increasing the number of turns, also enhances the accuracy.

Contents

List of Figures

List of Tables


Words That Count

Chapter 1

Introduction

Contents

Authorship analysis is the process of examining the characteristics of a piece of writing, a ancient text, a program code or comments on website etc. to draw conclusions on its authorship.

An important question arising when dealing with authorship analysis is whether the writing characteristics or style of an author evolves over time or changes with different contexts such as location, mood, time of day, presence of other people, etc. However, humans tend to have certain persistent personal traits. All hu-

mans also have unique patterns of behavior, much like the uniqueness of biometric data. Therefore, certain characteristics pertaining to language, composition, and writing, such as particular syntactic and structural layout traits, patterns of vocabulary usage, unusual language usage, and stylistic traits will remain relatively constant. The identification and learning of these characteristics with a sufficiently high accuracy is the principal challenge in authorship identification [30].

Based on some definitions from Gray et al.(1997), authorship analysis studies is categorized into three major fields: authorship identification, authorship characterization and authorship similarity detection.

1.1 Authorship Analysis

The first problem of authorship analysis is ***authorship identification*** [16, 19, 23, 28, 30, 43] that determines the likelihood of a piece of writing to be produced by a particular author by examining other writings by that author. It is also called "authorship attribution" in some literatures, especially by linguistic researchers. The problem can be considered as a statistical hypothesis test or a classification problem. The essence of this classification is identifying a set of features that remain relatively constant for a large number of writings created by the same person. Once a feature set has been chosen, a given written text can be represented by an n-dimensional vector, where n is the total number of features. Given a set of labelled vectors (i.e. a set of vectors with labels of the corresponding author), we can apply many analytical techniques to determine the category of a new vector extracted from a new piece of written text. Hence, the feature set and the analytical techniques may significantly affect the performance of authorship identification.

Formally, let $\{S_1, \ldots, S_n\}$ be a gallery set of possible authors of an anonymous text message ω. Let $\{M_1, \ldots, M_n\}$ be the sets of text messages previously written by the gallery $\{S_1, \ldots, S_n\}$, respectively. Assume the number of messages of each set of M_i, denoted by $|M_i|$, is reasonably large (say > 30). The problem of authorship identification is to identify the most plausible author S_a of ω from $\{S_1, \ldots, S_n\}$ with presentable and intuitive evidence. The most plausible author

S_a is the author whose writeprint of his text messages M_a has the "best match" with stylometric features in ω.

Authorship characterization [19, 23, 43] is the second problem of the authorship analysis. It summarizes the characteristics of an author of a given set of anonymous text messages and generates the author profile based on his or her writings. Some of these characteristics include gender, educational and cultural background, and language familiarity. Unlike previous problem that have training samples, there are no suspects and no training samples available for investigation in this problem. Thus, the problem of authorship characterization is how to infer the extracted behavioral and personal characteristics (form the writing style) of authors of the anonymous text messages by matching writeprints in the online text documents.

The third problem is ***authorship similarity detection*** [4, 13, 19, 23, 43] that compares multiple pieces of writing and determines whether they were produced by the same author without actually identifying the author. Each anonymous identity is compared to all other identities. Identities with a similarity score above a certain threshold are grouped together and considered to belong to the same entity (clustering task).
Most studies in this category are related to plagiarism detection. Plagiarism involves the complete or partial replication of a piece of work without permission of the original author. Figure 1.1 shows the difference between authorship identification and authorship similarity detection aspects.

1.1.1 Stylometry

Stylometry is defined as the "statistical analysis of writing style" [43].
Stylometry was born out of the need to identify authors of literary works and can be traced as far back as 1439 with Lorenzo Valla's work analyzing Donation of Constantine and his assertion that it was a forgery. Stylometry has successfully been applied to music and fine-art paintings as well.

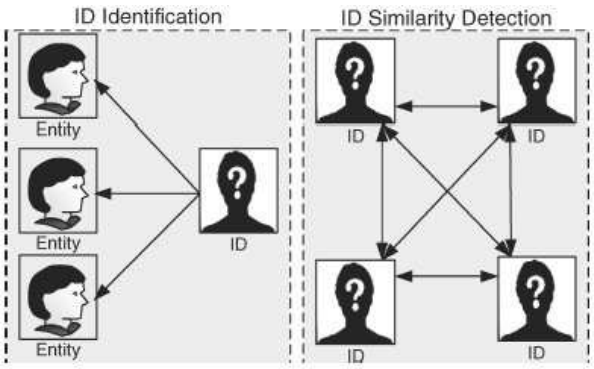


Figure 1.1: Purpose of authorship identification and authorship similarity detection

Over time stylometry has been used successfully in a number of different areas. The following list displays some of the uses of stylometry.

- Music lyrics [21]
- Music melody [21]
- Paintings [37]
- Literary works [10]
- Forensic Linguistics [1]
- Plagiarism [43]
- Social networking [23, 27]
- Electronic email [10, 27]
- Instant Messaging [1, 23]

AA aims at associating a portion of text with an author. It is based on a set of features, whose values identify with a certain degree of accuracy a persons. Techniques for extracting these features are many. Most of them comes from the implementation of some aesthetical rules of writing studied by a linguistic field named stylometry, which has old roots.
Based on previous studies, it is well established that feature sets can be broken out into five categories – Lexical, Syntax, Structural, and Content-specific and Idiosyncratic [4, 10, 43].

The vast majority of authorship attribution studies are based on ***Lexical features*** to represent the style. According to this family of measures, a text is viewed as

a mere sequence of characters or as a sequence of tokens grouped into sentences, where each token corresponds to a word, number, or a punctuation mark.
They can be further divided into character-based and word-based features. These can include sentence/line length [8], vocabulary richness which is calculated with a set of functions [13], and word length distributions [13, 43]. The vocabulary richness functions are attempts to quantify the diversity of the vocabulary of a text.
In [31] this family of features is defined as quantitative approach and it's considered the base of stylometry, which is centred on features that can be numerically counted and computed.
The most straightforward approach to represent texts is by vectors of word frequencies. That is, the text is considered as a set of words each one having a frequency of occurrence disregarding contextual information.

Syntactic features, including function words, punctuation, and part of speech, can capture an author's writing style. The discriminative power of syntactic features is derived from people's different habits of organizing sentences.
Syntactic information is considered more reliable authorial fingerprint in comparison to lexical information. Baayen, van Halteren, and Tweedie (1996) were the first to use syntactic information measures for authorship attribution.

Structural features represent the way an author organizes the layout of a piece of writing, paragraph indentation and signature-related features.
They are especially useful for online text, include attributes relating to text organization and layout [13, 43], technical features such as the use of various file extensions, fonts, sizes, and colors as in [2].
They include Internet slang that is a form of abbreviation for augmenting the information throughput and minimizing the time spent in writing. Today's Internet slang is said to have become mainstream with America Online's instant messenger program back in the early 1990's. Examples of Internet slang are BRB = be right back, TTYL = talk to you later, and LOL = laugh out loud.

In addition to "content-free" features such as frequency of function word, total number of punctuations, average sentence length, vocabulary richness, ***content-specific*** features are important discriminative features for online messages. The selection of such features is dependent on specific application domains.
On the Web, one user may often post online messages involving a relatively small range of topics whereas different users may distribute messages on different topics (the term topic states for the subject or theme of a speech, essay, thesis, discussion, conversation discourse).
For this reason, special words or characters closely related to specific topics may provide some clue about the identity of the author.

Idiosyncratic features include misspellings, grammatical mistakes, and other usage anomalies. Such features are extracted using spelling and grammar checking tools. Idiosyncrasies may also reflect deliberate author choices or cultural differences, such as use of the word "center" versus "centre".
In [31] this family is assessed as qualitative approach.

1.1.2 From literature to digital and Internet contents

Trying to determine the authorship of digital (email, social networking applications) or Internet content (emoticons, font color, font size, embedded images, hyper-links) presents some different and unique challenges that were not introduced with conventional stylometry.

Today, the ubiquity of the Internet and the millions of devices that people use to connect to the Internet, have generated a need to determine authorship of digital content.
This need originated because of the growth of *cybercrimes*. Cybercrimes can be defined broadly as criminal activity involving the use of information technology. With applications and communication methods like email, blogs, Facebook, Instant Messaging, Twitter, and on-line communications, exchanging of information across the globe has become easy and instantaneous. In many cases people that commit cybercrime are "hiding" behind the Internet anonymously or many times

use false identities.
Since text traces are often the only identity cues left behind in cyberspace, researchers have begun to use online stylometric analysis techniques as a forensic identification tool, with recent application to email [13], forums [43], and program code [22].

There are also many occasions in which we would like to identify the source of some piece of software. For example, if after an attack to a system by some software we are presented with a piece of it, we might want to identify its source. Typical examples of such software are Trojan horses, viruses, and logic bombs.

Also online messages, as the major channel of Web communication, are important sources for identity tracing in cyberspace. Compared with conventional targets of authorship identification such as literary works or published articles, one challenge of author identification of online and offline messages is the limited length of online messages. For example, email content tend be a lot shorter than some of the previously works on books. As Ledger and Merriam (1994) claimed, authorship characteristics would not be strongly apparent below 500 words. Based on their review of the size of writings in related studies, Forsyth and Holmes (1996) found that it was very difficult to attribute a text of less than 250 words to an author. The short length of online and offline messages may cause some identifying features in normal texts to be ineffective. For example, since the vocabulary used in short documents is usually limited and relatively unstable, measures such as vocabulary richness may be not as effective. Hence, how to correctly identify the authors of these relatively short documents with appropriate features becomes a challenge.

On the other hand, these kind of web messages also have some characteristics which may help reveal the writing style of the author. Since Web-based channels such as e-mail, newsgroup, and chat rooms are relatively casual compared with formal publications, authors are more likely to leave their own writeprints in their articles. For example, the structure or composition style used in online messages is often different from normal text documents, possibly because of the different purposes of these two kinds of writings. Most previous studies, as authorship

identification of Shakespeare's works and the Federalist Papers, dealt with a relatively small number of authors, typically no more than 10; and the average number of messages per author ranged from less than 10 to 300. Under these levels of parameter settings, satisfactory classification performance could be achieved. But in the context of identity tracing on the Internet, the number of potential authors for an online message could be large. Since cyber users often use different usernames on different Web channels, the number of available messages for each author may be limited. The brief of online messages present a challenge for authorship identification.

1.1.3 AA on chat: a new trend

Every second millions of Instant Messages (IM) are sent throughout the world employing heterogeneous applications such skype, twitter.
However there's a high probability that somebody enters into your account to send someone an IM on your name.

Stylometry may be able to assist in determining authorship of Instant Messages. IM's can be verified matching a particular user's stored chatting style.
Indeed it could be a part of the IM product to verify authenticity of the user automatically.

The study of AA on chat for the security of Web or defence from cybercrime is just a small portion of the reasons that have unearthed this area of research.
Other purposes for making AA on chat are the desire to discover the personality from writing style (John et al. 1991), to predict the age or gender as in [8, 20], to study deceivers beahvior [44], to determine an author's native language (Koppel et al. 2005).

After these considerations, we can assume that stylometry can be applied to IM conversations.

1.1.4 Instant Messaging

Instant Messaging (IM) is a popular form of computer-based communication. Since our study is focused in Instant Message conversations, we have to enlighten the difference with the other widely used computer-based communication called "chat". Whereas Instant Messaging is generally from individual to individual, and may have chat like features that allow multiple users to talk at the same time to each other, chat is usually an open "room" where several people may talk at once in a type of community. Chatting can covers chatrooms, IRC, and IM's.

By definition, IM is a communication service that enables its users to create a kind of private chat room with another individual that allows communication in real time over the Internet, similar to a telephone conversation but (typically) using text rather than voice.

There are dozens of Instant Messaging products available on the web. AIM (AOL Instant Messenger), Yahoo Messenger, Windows Live Messenger, Google Talk and Skype are some of the popular Instant Messaging applications. Interestingly enough, even with all the various networks being developed by corporations for profit, their physical structures (client-server architecture) and communication protocols (information packets) are very similar to one another.

Most Instant Messaging networks follow a strict Client-Server model in which a server (or a cluster of servers) is maintained by a service provider who controls traffic coming to and from the server. Users who wish to utilize a certain network generally register themselves with the service provider, then download a provider-approved client for using their network. Using this client, users can connect to the central server in order to be able to send and receive messages and collect account information.

A friend is generally another registered user. The concept is that a user may maintain a *Buddy List* with a list of its immediate friends and it may based upon the statuses of its friends.

Each member of the IM service has a flag , which indicates its status. The status informs the other people that one is ready to communicate, or that he/she does not want to do it. Possible user statuses are online, offline, idle, away, busy.

The Instant Messaging system alerts its users whenever somebody on their private list is online. Users can then initiate a chat session with that particular individual. IM technology lets users communicate across networks, in remote areas, and in a highly pervasive and ubiquitous manner.

Another important aspect of communication flow within an Instant Messaging network is the traffic of messages between users.
The IM conversations are recorded in a simple text format.

Because IM is a form of Computer-Mediated Communication, that closely resembles spoken interaction, IM is expressed through writing, but it share many of the characteristics of Face-to-Face (FtF) communication , but unlike FtF interaction, chat via IM is poor at managing interruptions, organizing turn-taking, conveying comprehension, and resolving floor control conflicts.

1.1.5 Social Signaling

There is now a growing research in cognitive sciences, which argues that our common view of intelligence is too narrow, ignoring a crucial range of abilities that matter immensely for how people do in life. This range of abilities is called social intelligence [6] and includes the ability to express and recognize social signals and social behaviours like turn-taking, agreement, politeness, and empathy, coupled with the ability to manage them in order to get along well with others while winning their cooperation.

Social intelligence is the facet of our cognitive abilities that aims at dealing effectively with social interactions and, at its core, it includes two main aspects. The first is the correct interpretation, in terms of social signals, of non-verbal behavioral cues displayed by others. The second is the generation of non-verbal cues

expressing social signals appropriate in a given situation.

Social signals and social behaviours are the expression of ones attitude towards social situation and interplay. They are manifested through a multiplicity of non-verbal behavioural cues including facial expressions, body postures and gestures, and vocal outbursts like laughter as in Figure 1.2.

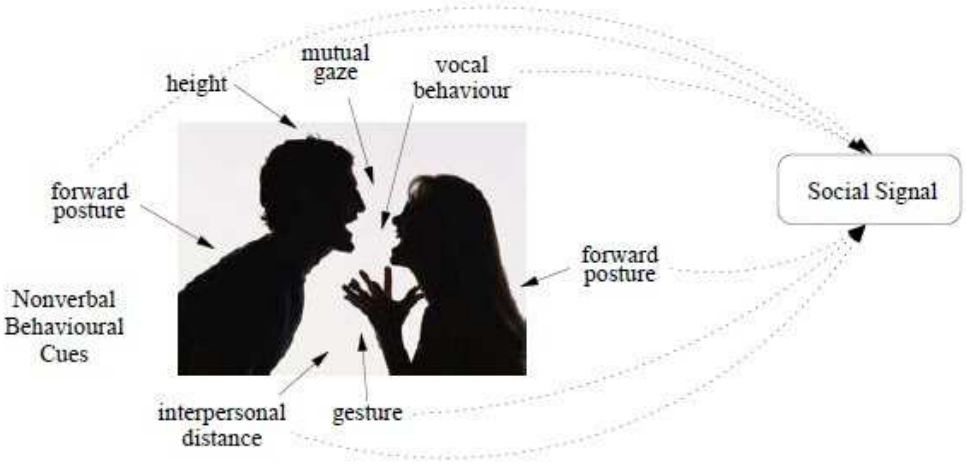


Figure 1.2: Behavioural cues and social signals

Social signals typically last for a short time (milliseconds, like turn taking, to minutes, like mirroring), compared to social behaviours that last longer (seconds, like agreement, to minutes, like politeness, to hours or days, like empathy) and are expressed as temporal patterns of non-verbal behavioural cues.

The research area of machine analysis and employment of human social signals to build more natural, flexible computing technology goes by the general name of Socially-Aware Computing as introduced by Pentland.

Researches in this area have coined the term Social Signal Processing (SSP) [41, 42], that is the new research and technological domain that aims at providing computers with the ability to sense and understand human social signals. The focus of Social Signal Processing is on non-verbal behavioral cues that human sciences (psychology, anthropology, sociology, etc.) have identified as conveying social sig-

nals.
In other words, SSP brings social intelligence in machines via modeling, analysis and synthesis of non-verbal behavior in social interactions. The rationale is that such cues are the physical, machine detectable and synthesizable evidence of phenomena nonotherwise accessible to computers such as empathy, roles, dominance, personality, (dis-)agreement, interest, etc.

1.1.5.1 Verbal and Non Verbal behavioural cues

The term behavioural cue is typically used to describe a set of temporal changes in neuromuscular and physiological activity that last for short intervals of time (milliseconds to minutes) in contrast to behaviours (e.g. social behaviours like politeness or empathy) that last on average longer (minutes to hours).

In most cases, behavioural cues accompany verbal communication and, even if they are invisible, i.e., they are sensed and interpreted outside conscious awareness, they have a major impact on the perception of verbal messages and social situations. Hence, this kind of behavioural cues are known as non-verbal behavioral cues.

For the sake of simplicity, psychologists have grouped all possible non-verbal behavioral cues occurring in social interactions into five major classes called codes.

- The first is ***physical appearance***, including not only somatic characteristics, but also clothes and ornaments that people use to modify their appearance.
- The second code relates to ***gestures and postures***, extensively investigated in human sciences because they are considered the most reliable cue revealing actual attitude of people towards others.
- ***Face and eye behavior*** is a crucial code, as face and eyes are our direct and naturally preeminent means of communicating and understanding somebody's affective state and intentions on the basis of the shown facial expression. Not surprisingly, facial expressions and gaze behavior have been extensively studied in both human sciences and technology.

- ***Vocal behavior*** is the code that accounts for how something is said and includes the following aspects of spoken communication: voice quality (prosodic features like pitch, energy and rhythm), linguistic vocalizations (expressions like "ehm", "ah", etc.) and non-linguistic vocalizations (laughter, crying, sobbing, etc.), silence (use of pauses), and turn-taking patterns (mechanisms regulating floor exchange). Each one of them relates to social signals that contribute to different aspects of the social perception of a message. In SSP the conversation analysis concerned for the most part on audio signals. Vocal behavior plays a role in expression of emotions, is a personality marker, and is used to display status and dominance. The speech analysis community has worked on the detection, e.g., of disfluencies, non-linguistic vocalizations (e.g., particular laughter), or rhythm, but with the goal of improving the speech recognition performance rather than analysing social behavior.

- The last code relates to ***space and environment***, i.e. the way people share and organize the space they have at disposition. People tend to organize the space around them in concentric zones accounting for different relationships they have with others.

With respect to vocal behavior, communication is a dynamic process with the interacting components of sending, receiving messages as stated by Park et al.(2009). As De Vito (2000) suggested, a message to have meaning, both elements, verbal and non-verbal, need to be present. The verbal indicators (described by the classical stylometry) are directly related to the spoken or written content, whereas non-verbal cues focus on accessory features that are exhibited while a person is producing content.

Non-verbal communication adds nuance or richness of meaning that cannot be communicated by verbal elements alone. De Vito defines non-verbal communication as communicating without words: "You communicate non-verbally when you gesture, smile or frown, widen your eyes, move your chair closer to someone wear jewellery, touch someone, raise your vocal volume, or even when you is nothing". Hence, non-verbal communication regulates relationships between messages and meaning, and may support or replace verbal communication with non-verbal be-

haviors. The comparisons of verbal and non-verbal communication are as follows:

- An adult interprets messages more depending on non-verbal information compared with verbal one.
- Children interpret messages more depending on verbal information.
- Verbal communication is used to transfer the actual, concise, and reliable messages, while non-verbal communication is used to provide accuracy, clarity, and contradiction for ambiguous messages.

The properties of non-verbal communication can be also found in the neurophysiology.

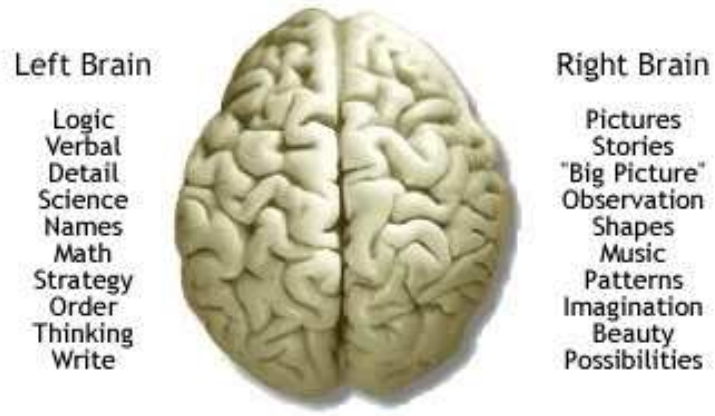


Figure 1.3: The functions of a right brain and a left brain

Extracting social information from non-verbal communication is hard wired in the human brain. Any facial expression, vocal outburst, gesture or posture triggers often unconscious analysis of socially relevant information. Furthermore, this mechanism seems to be so deeply rooted in our brain, that we cannot escape it, even when we deal with synthetic faces and voices generated by computers.
In the view of non-verbal information as the neurophysiology, a left brain operates the analytic, logical, linguistic process of thinking, while a right brain operates the emotional, intuitive process of thinking as shown in Figure 1.3.

Non-verbal communication is the wide spectrum of non-verbal behavioral cues that we display when we interact with others, with machines and with media. From

a computing point of view, this is important for two reasons. The first is that non-verbal behavioral cues play the role of a physical, hence machine detectable evidence of social signals. The second is that non-verbal cues synthesized through some form of embodiment (conversational agents, robots, etc.) express the same relational attitudes as when they are displayed by humans, thus are likely to synthesize social signals.

This is exactly the problem addressed by Social Signal Processing (SSP). Indeed, the core idea of SSP is that non-verbal behavioral cues can be detected with microphones, cameras, and any other suitable sensors.

1.1.5.2 Non-verbal information in Computer-Mediated Communication

Computer-Mediated Communication is a process of human communication via computers, involving people, situated in particular contexts, engaging in processes to shape media for a variety of purposes.
This includes communication both to and through a personal or a mainframe computer, and is generally understood to include asynchronous communication via email or through use of an electronic bulletin board, synchronous communication such as "chatting" or through the use of group software, and information manipulation, retrieval and storage through computers and electronic databases.

Computer-Mediated Communication becomes more popular communication form, and replaces the traditional face-to-face communication fields gradually. In Computer-Mediated Communication, there consists of four major parts as follows in Table 1.1

	Synchronism	**Purpose**	**Style**	**Transmission of Message**
E-mail	Asynchronism	Information exchange	Text	One to one One to many
E-bulletin board	Asynchronism	Information exchange	Text, discussion	One to many
Chat	Synchronism	Social interaction	Text	One to server
MUD	Synchronism	Social interaction	Text, graphics	One to server

Table 1.1: Examples of Computer-Mediated Communication

Since the human communication is strongly supported by non-verbal signals, their

absence results in a poor exchange.
To solve the problem, several approaches have been drawn but the success with which non-verbal cues are modeled is low, mostly because it reduces in a artificial, unconventional use of the verbal cues. The emoticons are an example of such a mechanism: they represent differents mood (happy, sad, angry, etc) by means of code letters.

By taking unconventional cues such as response latency and keystroke activities as a special type of paralinguistic behaviour (vocal and sometimes non-vocal signals beyond the basic verbal message or speech, includes pitch, loudness, rate, and fluency), Gajadhar and Green extended the scope of paralinguistic behaviour beyond the traditional definition.
Example of such paralinguistic behavior are: exclamation for emphasis, show happiness, question, show agreement, negative emotion, negative exclamation, exit word, emphasis, positive exclamation.

1.2 Goals

The primary goal of the thesis is that of performing Authorship Attribution on written texts of IMs. In specific, dyadic conversations will be taken into account, and in particular Skype chats. For this aim, our idea is that of exploiting the relationship that Instant Message communication shares with the spoken communication, i.e., the fact of being structured with temporized turns. Such a connection will be employed to extract non-verbal cues for characterizing the profile of a person, while she is chatting. In turns, this will be used to perform authorship attribution in a unconventional way.

There were very few scientists that do this stuff, and most of them considered chats as ordinary text.
The experiments are performed over a corpus of dyadic chat conversations that involves 77 subjects.

The rest of the work is organized as follows:

Chapter: State of Art shows a review of the previous works. It considers the evolution of stylometry over the years, showing that the first application were in simply text categorization or classification. Then, with the advent of computers, it has been possible to use advanced tools to build AA techniques. In the last years, the main research trend has become the chat analysis.

Chapter: Mathematical Background explains the techniques and measures that we used in the experiments.

Chapter: Method first shows how Skype client works. Then our feature extraction from raw data is applied. This extraction contains the verbal and non-verbal features that concerns turn-taking conversation.

Chapter: Experiments and results obtained after a feature selection process.

Conclusion reports comments of this study, problems encountered and possible future works.

Chapter 2

Related Works

Contents

In this chapter a summary of a taxonomy and review of stylometric analysis research is presented.

2.1 The evolution of stylometry

A significant amount of research has been done in the area of authorship attribution. The use of stylometry and authorship identification to determine authorship dates to precomputed times [10].

Stylometry was born out of the need to identify authors of literary works and can be traced as far back as 1439 with Lorenzo Valla's work analyzing Donation of Constantine and his assertion that it was a forgery. Stylometry has successfully been applied to music and fine-art paintings as well.

The investigation of authorship attribution has existed for centuries. For example, in the 1700's, Edmond Malone questioned whether or not Shakespeare [40] really wrote some of the plays bearing his name.

The origins of term of stylometry date back to 1851 when the English logician Augustus de Morgan suggested in a letter to a friend that questions of authorship might be settled by determining if one text "does not deal in longer words" than another (de Morgan, 1882) [18]. His hypothesis was investigated in the 1880's by Thomas Mendenhall, an American physicist who subsequently published the results of an impressive display of academic labour on word length "spectra" (Mendenhall, 1887) [39]. Since then stylometrists have been searching for a unit of counting which enumerates the style of the text. The application of counting the features of a text was later extended by Yule in the early 1900's to include the length of sentences.

Williams (1940) discovered that by charting frequency distributions of the logarithms of the number of words per sentence, an approximation to a normal distribution was obtained for each author, a finding later used by Wake (1957) in his study of Greek authors. This was a long, painstaking, manual process centuries ago however with the advent of the computer, automated methods for feature selection and authorship attribution have proven to be quite successful.

Stylometry primarily concerns itself with attribution studies, although chronological studies on the dating of work within the corpus of an author have also been investigated (Cox and Brandwood, 1959).
The most thorough and convincing study in this field was conducted by Mosteller and Wallace (1964). In their study on the mystery of the authorship of the Federalist Papers, they attributed all 12 disputed papers to Madison. Their conclusion was generally accepted by historical scholars and became a milestone in this research field. They first used the frequency of occurrence of function words (e.g., "while" and "upon") to clarify the disputed work.

Baley (1979) lists the general proprieties which quantifiable feature of text should

possess: "they should be salient, structural, frequent, easily quantifiable and relatively immune from conscious control". By measuring and counting these features, stylometrists hope to uncover the "characteristics" of an author. An author's linguistic style is thought to have certain features that are independent of the author's will, and since these features cannot be consciously manipulated by him, they are considered to provide the most reliable data for stylometric study.

Morton's technique was applied by Merriam in three studies involving Shakespeare (1979, 1980, 1982), yet the technique was seized upon by Smith (1985) who condemned it for lack of rigour, dubious data acquisition, and the small size of samples used.
The Smith versus Morton arguments spilled over into the use of once-occurring words (hapax legomena) as discriminatory features.
In the late 1980s and the early 1990s, Burrows use large sets of common high-frequency function words, compute their rates of occurrence per thousand words in the candidate texts, and employ what is essentially principal components analysis on the resulting multivariate data array.

Since then and until the late 1990s, research in authorship attribution was dominated by attempts to define features for quantifying writing style, a line of research known as "stylometry".
Cumulative sum charts or "cusum" charts are statistical techniques primarily used in industrial processes and quality control monitoring. In the early 1990s, Morton proposed an authorship test which used cusum charts, and his ideas were put forward in two internally published reports (Morton and Michaelson, 1990; Morton, 1991).
The style can be tremendously affected by demographic factors, including gender and age of the writer. Of central importance was Morton's claim that each person has a unique set of habits which he or she follows consistently whenever communicating, whether through the written or spoken word. These habits are quantifiable in that they are particular components of that person's sentences; those relied upon "short words" (defined as words of two or three letters), "vowel words" (defined as words beginning with a vowel), and especially the combination

"short + vowel words".
The cusum controversy did not deflect stylometry from its evolutionary path, and its modern face has now been changed by the influx of techniques from the domains of computer science and artificial intelligence.

In 1995 Krusl and Spafford [35] explored the classification of programme's style, and tried to find a set of characteristics that remain constant for a significant portion of the program that a programmer might produce. Using "Software Metrics" they analysed C source code [22].
A useful introduction to neural network applications has been provided by Tweedie et al. [39]. The initial work involving neural networks with stylometry was presented for Shakespeare/Fletcher controversy.
Baayen et al. (1996) have compared the performance of various stylometric techniques when fed with different kind of features, in specific the lexical vocabulary and the syntactic. They conclude that the more syntactically aware methods give better classificatory accuracy, but they noted that the application of such techniques would require the existence of many more syntactically annotated corpora.
Holmes (1998) analyzed the use of "shorter" words (i.e., two- or three-letter words) and "vowel words" (i.e., words beginning with a vowel). Such word-based and character-based features require intensive efforts in selecting the most appropriate set of words that best distinguish a given set of authors and sometimes those features are not reliable discriminators when applied to a wide range of applications.

Koppel et al. in 2002 try to categorize written text by author gender [20]. They considered categorization by author style, and intonational, phonological, and conversational cues. Their objective is to use a set of training documents to find a linear separator between male and female documents. They use a variant of the Exponential Gradient and Balanced Winnow algorithm. They also apply a feature reduction and for each model they obtained in a cross-validation trial.

McCarthy et al. [26] use Coh-Metrix, a computational tool that analyzes text on over 200 indices of cohesion and difficulty. They demonstrate how, with the benefit of statistical analysis, texts can be analyzed for subtle, yet meaningful differences.

They report evidence that unlike genres and modes, however, an author's style can vary over time. This instability leads to a problem for computational identification through style markers. The problem, is that researchers expect a textual feature to be both static enough (to distinguish one author's works from another's), but at the same time, variable enough (to indicate where in the author's career an undated text may fit).

Park et al. in 2009 analyze writing stiles of bloggers with different opinions. From the corpus of blog they extracted lexical feature, structural feature as organization of content through the use of sentences and paragraphs and the sentimental features indicate the direction and degree of sentiment expressed by the words in the text. After extracting the stylometric features from each blog entry, they calculated their mean values for each author.
Shalhoub et al. in 2010 [34] discuss potential uses of stylometry in the area of Internet content and present four use cases in areas that have had little or no research: social network, email, chat and terrorism use cases.

With the advent of Social Networks [16, 33, 34, 44], the research on interaction and relationship between Facebook and Twitter users grown.
Goldbeck et al. [15] tried to predict personality form Twitter posts and public information on user's Twitter profile. They considered the "Big Five" model to classify personality. The models five domains of personality, openness, conscientiousness, extroversion, ageeableness, and neuroticism, were conceived by Tupes and Christal[38].
After administering 45-questions version of Big Five Personality Inventory -to the fifty subject recruited through posts on Twitter, Facebook and relevant mailing lists - they collected 2000 tweets from users.
For each one, they collect number of followers, followings, mentions, replies, hashtags, link and words per tweet. For the number of mentions, replies, hashtags, and links, they used the raw numbers and the average per tweet. Then, they merged all tweets in a single document for a given user. They included words count, words per sentence, and swear word counts since these reflect verbosity and tone of the user.

For the other categories, the values are given as the percentage of words in the input that match words in a given category. They compute average and standard deviation score on each personality factor on a normalized 0-1 scale. To predict the score of a given personality feature they performed a regression analysis on Weka using Guassian Process and ZeorR each with 10-fold cross-validation with 10 interactions.

In 2011 Castro et al. used False Acceptance Rate (FAR) and False Rejection Rate (FRR) to show that Keystroke System is functional and sure, and provide an indication of usefulness of stylometry for identify the online test takers.
This was accomplished by using a combination of the keystroke and stylometry features. The resulting output displays the type of tests that were run, sample test size of intra class and inter class, test size, FRR, FAR, test subject samples averages and KNN.
Their stylistic features include lexical, syntactic, structural, content-specific, and idiosyncratic style markers.
FAR is the measure of the likelihood that the biometric security system will incorrectly accept as access to attempt by an unauthorized user, that is the ratio of the number of false acceptance divided by the number of identification attempts.
FRR is the measure of the likelihood that the biometric system fails to verify an authorized user, that is the ratio of the number of false rejection divided by the number of identification attempts [11].

2.2 Stylometry and Authorship analysis works

In this section is presented a review of authorship attribution methods.

O. de Vel et al. [13] investigated into email content mining for forensics authorship identification or attribution.
They focused their discussion on the ability to discriminate authors in the case of both aggregated e-mail topics as well as across different email topics. They wish to investigate the degree of orthogonality existing between e-mail authorship and e-mail topic content.

An extended set of e-mail document features including structural characteristics and linguistic patterns were derived and, together with a Support Vector Machine learning algorithm, were used for mining the e-mail content.
In the same year they investigated the gender and language background of an author, based on cohort attribution mining from e-mail text documents. In addition to some well known features, they use features like smiles and emoticons.

In 2003 Corney [12] advocated that stylometry is a valuable tool for computer forensics and investigation to determine authorship of anonymous messages. By analyzing email texts he came to the conclusion that a combination of character based, word length frequency distribution, and function word attribute is an effective combination of features.
Gamon in 2004 [14] studied authorship classification demonstrating that a combination of features based on shallow linguistic analysis (function word frequencies, part of speech trigrams) and a set of deep linguistic analysis features (context free grammar production frequencies and features derived from semantic graphs) yields very high accuracy in attributing a short random text sample to one of the three Brontë sisters as its author.
In order to be maximally "content-independent", they normalized all personal pronouns to an artificial form in order to not pick up on "she" or "he" frequencies which would be linked to the gender of characters in the works of fiction rather than author style.
He shown that the use of deep linguistic analysis features in authorship attribution can yield a significant reduction in error rate over the use of shallow linguistic features such as function word frequencies and part of speech trigrams.

Goodman et al. in [17] make an attempt to explain the methodology of developing and optimizing a stylometry program which quantifies the writing styles of various authors using 62 stylistic features found in emails.
Raw keystroke data, collected from an Internet-based Java applet in an earlier keystroke biometric study, was converted into simple text files, appropriate features were extracted and a pattern classifier was implemented.
For a comparative analysis, the raw frequency counts of each stylistic feature are

normalized from a 0 to 1 range and classified by the k-nearest neighbor algorithm using Euclidean distance. Two products were to be produced from the extraction: a reconstruction of the original text and a "dirty" file that contained every keystroke entered.
Any non-printing keystrokes - such as Shift, Alt- were to also be included in the dirty file.

In 2008 Calix et al. work to optimized and extend an existing C# based stylometry system developed by Goodman et al., that identifies the author of an arbitrary e-mail by using fifty-five writing style features. This program uses the K-Nearest Neighbor algorithm [10].
Goldstein et al. in 2009 [16] conducted an experiment on identifying a person and identifying e a person by genre using text sample including essays, emails, chat, blogs and the audio samples were transcribed to text.
They considered six genres on six topics, and from the corpus of text they extracted feature by LIWC, function words, specific topic words, stop word, abbreviation, emotional expression and emoticons. They use eight feature set combinations for the classification. They use cross validation across genres to identify a person ad an author of a text.

In Argamon [8] unlike the problem of authorship attribution, authorship profiling does not begin with a set of writing samples from known candidate authors.
Instead, he exploits the sociolinguistic observation that different groups of people speaking or writing in a particular genre and in a particular language use that language differently. example).
There are two basic types of features that can be used for authorship profiling: content-based features and style-based features.
Many different types of features have been considered as possible markers of textual style including lexical, syntactic, and vocabulary complexity-based features.
They use as learning algorithm Bayesian Multinomial Regression (BMR) which they have found to be both efficient and accurate. BMR is a probabilistically well-founded multivariate variant of logistic regression which is resistant to overfitting. They consider four profiling problems: determining the author's gender,

age, native language, and neuroticism level.
They also emphasized difficulties in identifying authorship of an e-mail text for two of the following reasons: firstly, the concise nature of e-mail messages (tens or perhaps hundreds of words comparing to thousands for articles and books) and secondly, the variation in the individual style of e-mail messages due to the fact that e-mails, as an informal and fast-paced medium, exhibit variations in an individual's writing styles due to the adaptation to distinct contexts or correspondents.

Narayanan et al. in 2012 try to identifying an anonymous author of blogs text by linguistic stylometry. They use a very huge database of 100,000 candidate author to demonstrate authorship recognition using both "lazy " and "eager" classifier, such as nearest neighbor (NN), naive Bayes (NB), support vector machine (SVM) and regularized least squares classification (RLSC). Furthermore, they try to motivate the development of completely automated tools for transforming one's writing style while preserving the meaning [28].
The work focuses on style and behavior; the respective deanonymization algorithms show a natural progression in complexity as well as in the amount of data required. In the experiments they test classifiers which return a ranking of classes by likelihood, rather than those which can only return the most likely class.
They do make use of single character frequencies, excluding bigrams and trigrams, which may be significantly influenced by specific words. Stanford Parser is used determine the syntactic structure of each of the sentences in the input posts.
As output, it produces a tree for each sentence where the leaf nodes are words and punctuation used, and other nodes represent various types of syntactic categories (phrasal categories and parts of speech). They generate features from the parse trees by taking each pair of syntactic categories that can appear as parent and child nodes in a parse tree tree, and counting the frequency of each such pair in the input data. They compute information gain for each feature in the entire dataset to understanding better the meaning of each feature and a confidence estimator to obtain confidence score, on different threshold, to achieve various trade-off between precision and recall.
This work has a few important limitations. First, the attack is unlikely to work if the victim intentionally obfuscates their writing style. Second, while they have

validated our attack in a cross-context setting (i.e., two different blogs), they have not tested it in a cross-domain setting (e.g., labelled text is a blog, whereas anonymous text is is an e-mail).

In the follows is collected the major works, splitted by used classifier in Table 2.1, metrics in Table 2.2 and features in Table 2.3.

Study	**Classifier**
Krusl and Spafford (1995)	software metrics, SAS tool
koppel et al.(2001)	variant of exponential gradient and Balanced Winnow algorithm
De Vel et al. (2001)	SVM
Gajadhar and Green(2003)	not used
Gamon (2004)	SVM
Resig and Teredesai (2004)	clustering
Resig et al. (2004)	clustering on status
Zhou and Zhang (2004)	not used
McCarthy et al.(2006)	coh-matrix
Orebaugh (2006)	k-NN
Zheng (2006)	C4.5, NN, SVM
Goodman et al.(2007)	k-NN
Abbasi et al.(2008)	techniques not classifier (pca, markov model, N-gram Models, Cross entropy, K-L similarity, Writeprint
Calix et al. (2008)	k-NN
Kukukylmaz et al(2008)	SVM,k-NN, PRIM, NB
Goldstein et al.(2009)	Random forest, NB, SMO, J48
Argamon et al. (2009)	BMR
Orebaugh and Allnut (2009)	j48 decision tree, IBK-NN, NB
Goswami et al.(2009)	NB
Park et al.(2009)	note used
Iqbal et al.(2011)	END, j48, RBFNetwork, NB, Baysnet, AuthorMiner, AuthorMIner2
Castro et al. (2011)	k-NN
Ali et al. (2011)	JW cross enropy, KS distance, Camberra distance, cosine distance, histogram distance, manhattan distance, Kullback Leibler distance, Levenshtein distance, intesection distance, LDA, RN cross entropy,Naive Bayes classifier, LZW distance, Mean distance
Narayanan et al. (2012)	NN, LDA, NB, Binary classifier, SVM, RLSC

Table 2.1: Summary of the different classifiers employed for AA

Study	Type	Features
Mendenhall (1887)	text	character frequency
Monsteller and Wallace(1964)	text	function word
Krusl and Spafford	code	programming layout metrics(indentation, separator, commenting style), programming style metrics (mean program line length, name length, naming conventions. conditional compilation, preference of while, for or do loops, use of comments), programming structure metrics(use of int, void, function, error detection, precentage of global variable, quality of software)

Table 2.3: continues on next page

Table 2.3: continued from previous page

Study	Type	Features
Koppel et al.(2001)	text	function words, n-grams of part of speech (preposition, singular noun, article), punctuation mark
De Vel et al. (2001)	email	number of blank lines/total number of lines, average sentence length, average,word length, vocabulary richness, total number of function word/total number of word, function word frequency distribution, total number of short words/total number of word, count of hapax legomena/total number of word, count of hapax legomena/total number of distinct word, total number of characters in words/total number of characters in body email, total number of alphabetic characters in words/total number of characters in body email, total number of uppercase characters in words/total number of characters in body email, total number of digit characters in words/total number of characters in body email, total number of white-space characters /total number of characters in body email, total number of space characters/total number of characters in body email, total number of space characters in words/ number white-space characters, total number of tab/total number of characters in body email, total number of tab spaces/number white-space characters, total number of punctations/total number of characters in body email, word length frequency distribution/total number of word, has greeting, use a farewell, contains signature text, number of attachment, position of requoted text within email body, HTML tag frequency distribution/total number of HTML tags
Gaiadhar and Green(2003)	chat	multi ..., multi !!!, multi ???, capitals, LOL, See ya, okay, :-), oops, oh, yep, wow, hey, exclamation for emphasis, show happiness, show agreement, question, negative exclamation, exit word, negative emotion, emphasis, positive exclamation
Resig and Teredesai (2004)	chat	user status, chat messages
Gamon	text	average length of sentences, noun phrases, adjctivial/adverbial phrases, frequencies of function words, frequencies of POS trigrams, frequencies of context-free grammar production, semantic features (number and person features on nouns and pronouns, tense and aspectual features on verbs, and subcategorization features on verbs, n-gram frequencies
Resig et al. (2004)	chat	user status, total online second and number of time that user change status, count the number of seconds x and y are online at the same time
Zhou and Zhang (2004)	chat	Productivity:total number of words composed by a participant in the entire conversation Participation:number of turns,average pause intervals between 2 messages,response latency referring to the average delay for a participant to send put a response Spontaneous correction measured by the ratio of erased messages
McCarthy et al.(2006)	text	Causal Cohesion, Coreferential Cohesion, Connectives and Logicial Operators,Density of Major Parts of Speech,Polysemy and Hypernymy,Syntactic Complexity,Word Information and Frequency

Table 2.3: continues on next page

Table 2.3: continued from previous page

Study	Type	Features
Orebaugh (2006)	chat	uppercase, lowercase, special characters and numbers as frequency distribution, word frequency distribution, emoticon frequency distribution, function word distribution, short word frequency distribution, punctuation frequency distribution,average word length, average words per sentence, contains greeting, contains farewell, abbreviation frequency distribution, spelling error, grammatical errors
Zheng et al.(2006)	misc	lexical, syntactic, structural, content specific, idiosyncratic
Goodman et al.(2007)	e-mail	number of sentences per paragraph, average word length, number of words, paragraphs and average number of words per paragraph, average number of paragraphs, average number of sentences, average number of words, average number of words per sentence, average number of white space, average number of commas, average number of periods, Number of Accents, Number of Left curly braces, Number of Right curly braces, Number of Vertical lines, Number of Tildes, Number of Windows keys, Number of Up keys, Number of Left Shift keys, Number of Right Shift keys, Number of Page Down keys, Number of Insert keys, Number of Home keys, Number of End keys, Number of Down keys, Number of Ctrl keys, Number of Context menu keys, Number of Caps Lock keys, Number of Alt keys, Number of F12 keys, Number of Right keys, Number of Backspace keys, Number of Enter keys, Number of Delete keys, Number of Tab keys, Number of words, Number of sentences, Average words per sentence, Number of paragraphs, Average words per paragraph, Average word length, ,Number of sentences beginning with lower case, Number of White spaces, Number of exclamation points, Number of Number signs, Number of Dollar signs, Number of percent signs, Number of Ampersands, Number of Single quotes, Number of Left parentheses, Number of Right parentheses, Number of Asterisks, Number of Plus signs, Number of Commas, Number of Dashes,Number of Periods, Number of Forward slashes, Number of Colons, Number of Semi-colons, Number of Less than signs, Number of Equal signs, Number of Greater than signs, Number of Question marks, Number of multiple question marks, Number of multiple exclamation marks, Number of ellipsis, Number of At signs, Number of Left square brackets, Number of Back slashes, Number of Right square brackets, Number of Carot signs, Number of Underscores, Number of sentences beginning with upper case
Abbasi et al.(2008)	comment	lexical, syntactic, structural, content specific, idiosyncratic

Table 2.3: continues on next page

Table 2.3: continued from previous page

Study	Type	Features
Calix et al. (2008)	email	Number of sentences beginning with upper case, Number of sentences beginning with lower case, Number of Words, Average Word Length, Number of Sentences, Average Number of Words per Sentence, Number of Paragraphs, Average Number of words per Paragraph, Number of Exclamation Marks, Number of Number Signs, Number of Dollar Signs, Number of Ampersands, Number of Percent Signs, Number of Apostrophes, Number of Left parentheses, Number of Right parentheses, Number of Asterisks, Number of Plus Signs, Number of Commas, Number of Dashes, Number of Periods, Number of Forward Slashes, Number of Colons, Number of Semi-colons, Number of Pipe Signs, Number of Less than Signs, Number of Greater than Signs, Number of Equal Signs, Number of Question Marks, Number of At Signs, Number of Left square brackets, Number of Right square brackets, Number of Backward slashes, Number of Caret Signs, Number of Underscores, Number of Accents, Number of Left curly ,Number of Right curly braces, Number of Ellipsis, Average Number of Periods per Paragraph, Average Number of Commas per Paragrap, Average Number of Colons per Paragraph, Average Number of Semi-colons per Paragraph, Average Number of Question Marks per Paragraph, Average Number of Multiple Questions Marks per Paragraph, Average Number of White Spaces per Sentence, Number of times "Well" appears, Number of times "Anyhow" appears, Average Number of Times the word "Anyhow" appears, Average Number of Times the word "Well" appears
Kukukylmaz et al(2008)	chat	character frequency, average message length, average word length, punctuation mark frequency, stop word frequency, smiles frequency, number of distinct word
Argamon et al.	text	content-based, style-based
Goldstein et al.(2009)	misc	stop words, frequencies of words, five topic specific words excluded from stop words, abbreviations, emotional expressions, emoticons, function words
Orebaugh and Allnut (2009)	chat	character frequency, emoticons frequency, word frequency,short word frequency, function word frequency, punctuation frequency, average word length, average word per sentence, contains a greeting, contains a farewell, abbreviation frequency, spelling error, grammatical errors
Goswami et al.(2009)	blog	sentence length, non dictionary words (smiley, slang, out of dictionary word, chat abbreviation
Park et al.(2009)	blog	lexical feature as number of words, average word length, standard deviation of word length, number of characters, frequency of upper-case, lowercase, special and numerical characters and eleven word-based features, as frequency of hapax legomena (i.e., words that occur exactly once in an entry), function words, structural feature as organization of content through the use of sentences and paragraphs (as total number of sentences and average number of words per sentence) and the sentimental features

Table 2.3: continues on next page

Table 2.3: continued from previous page

Study	Type	Features
Castro et al. (2011)	text	number of alphabetic character/total number of characters, number of uppercase character/total number of alphabetic character, number of digit/total number of character, number of space character/total number of character, number of vowel/number of alphabetic character, number of alphabetic char(a,e,i,o,u,upper or lowercase)/number of vowel character for each vowel, number of most frequent consonant character(t,n,s,r,h)/number of alphabetic character, for each consonant number of alphabetic consonant/number of most frequent consonant, number of least frequent consonant/number of alphabetic character, number of consonant-consonant digrams/total number of alphabetic letter digrams, number of consonant-vowel/total number of alphabetic letter digrams, number of vowel-vowel/number of alphabetic letter digrams, for full list see paper
Iqbal et al.(2011)	misc	lexical, syntactic, structural, content specific, idyosincratic
Ali et al. (2011)	chat bot	vowel 2-3 letters words, vowel 2-4 letters words, 2-4 letters, vowel inital words, 2-3 letters, character bigrams, characters, character trigrams, words, character tetragrams, 3-4 letters, MW function words, word bigrams, vowel 3-4 letters word, word lenght, syllables per word, hapax-dis legomena, word tetragrams, hapax legomena, dis legomena
Narayanan et al. (2012)	blogs	number of words/character in post, vocabulary richness, frequency of word with different combination of upper ad lowercase letters, frequency of word that have 1-20 characters, frequency of a to z ignoring case, frequency of 0 to 9, frequency of punctuation marks, frequency of other special characters, frequency of words like 'the', 'of', 'then', frequency of every pair (A,B) where A is the parent of B in the parse tree

Table 2.3: Summary of extracted features for AA analysis

Table 2.3: ends from previous page

2.3 Stylometry on Chat works

Thus, it is a natural extension to apply the techniques used for e-mail, forensics, and other purposes to IM author profiling and identification.
Gajadhar and Green analyse students interaction in online chat, in order to determine the form of any text-based non verbal communication used by them to enhance the meaning of their messages.
They determine that punctuation and typographical symbols have been used to display emotional cues that are missing in online chat.
For example, "Mmmmmmmmm" is a spoken pause used to show thinking, uncertainty, or agreement.

Study	Performance metric
Krusl and Spafford (1995)	metrics PRO1M, mean
Koppel et al.(2001)	accuracy from cross validation
De Vel et al. (2001)	accuracy, recall, precision, macro average F=2RP/R+P
Gajadhar and Green(2003)	not used
Gamon(2004)	accuracy, precision, recall, F-measure
Resig and Teredesai (2004)	level of confidence of user activities
Resig et al. (2004)	accuracy, similarity is assumed to exist between every pair of members in the same cluster, score as function of the probability of a pari of user interacting with each other
Zhou and Zhang (2004)	not used
McCarthy et al.(2006)	accuracy, precision, recall, F1-measure
Orebaugh (2006)	euclidian distance as distance metric, degree of confidence
Zheng et al. (2006)	accuracy
Goodman et al.(2007)	accuracy
Abbasi et al(2008)	F-measure, p-values
Calix et al. (2008)	accuracy
Kukukylmaz et al.(2008)	chi-square,p-value, z-score
Argamon et al. (2009)	not used
Goldstein et al.(2009)	accuracy
Orebaugh and Allnut (2009)	accuracy, error, true positive, false positive rate,
Goswami et al.(2009)	accuracy, confusion matrix
Park et al.(2009)	not used
Iqbal et al.(2011)	accuracy, weighted accuracy, sum
Castro et al. (2011)	accuracy by False Acceptance Rate (FAR-is the measure of the likelihood that the biometric security system will incorrectly accept as access to attempt by an unauthorized user, that is the ratio of the number of false acceptance divided by the number of identification attempts) and False Rejection Rate (FRR-is the measure of the likelihood that the biometric system fails to verify an authorized user, that is the ratio of the number of false rejection divided by the number of identification attempts)
Ali et al(2011)	accuracy
Narayanan et al. (2012)	threshold applied to score of confidence estimator to obtain a trade-off between precision and accuracy, gap statistic

Table 2.2: Summary of the different performance metrics employed for AA

The use of emoticon, smiley face :-), denote a friendly person and encourage friendly discourse. An acronym, "LOL" (laughing out loud), is used to signify laughter. Incomplete sentences, and misspellings may indicate the speed of the exchanges. "lo" used instead of "hello" may illustrate not only the speed but also the informality of the genre.

The use of the contraction is indicative of a conversational style and structure, but its repetition is indicative of support building. Question marks and repetition of "h" show emphasis or it means "I am not sure what you meant".

Emotive word, sorry, repetition of full stops for pause and effect, hesitation, slang, repetitions of sounds, for emphasis or admission of mistake misspelling, are all non-verbal cues. Also capitals, abbreviations, and exclamations such as oops,

"mmmmm" to signify feelings. Frequent use of onomatopoeia, for example, words such as whizz, eek, eh, aaarrr, (suggesting relief, frustration, annoyance, or humour) or beep beep beep (a desire to get into the conversation, meaning "Make way, I'm coming" or "Let's move on").
The repetition of these sound words is open to a variety of interpretations, or possibly misinterpretations. So, non-verbal communication is used to provide accuracy, clarity, and contradiction for ambiguous messages.
Once each item was manually noted and totalled on each page of the log, the totals per log were recorded and the items were then sorted by the total of the number of times of occurrence of the item over the period to determine which items were used most frequently.
The most non-verbal expressions used, were estimated to be multi special characters as "!!!!", "??????", "...", capitals, "LOL", emoticons, "ops", "oh","Hey", and "Wow".

Zhou and Zhang [44] explore online behavior of deception in a group IM setting, Yahoo more specifically.
They measured the "productivity", "participation" and "spontaneous correction" based on the role of the person in the discussion task.
As dependent variables they use "productivity" that is the total number of words composed by a participant in the entire conversation; "participation" that is measured by the number of turns a participant takes to exchange messages, the average pause intervals between sending two messages, response latency referring to the average delay for a participant to send out a response, initiation that indicate whether a participant initiated the discussion and "spontaneous correction" that is measured by the ratio of erased messages that are completely deleted first and then replaced with new ones.
The only independent variable is role played by participants. The chat conference was set up in such a way that all messages were archived and time-stamped. Each conference consisted of three participants along with a facilitator from the research team. Thirty-six undergraduate student volunteers were recruited for tests and they show the mean and standard deviation for individual dependent measures [44].

In 2004 IM clients are analysed by Resig and Teredesai: AOL, Yahoo, MSN. They explore data mining issues and how they relate to IM and current counter-terrorism efforts [33].
The Study focus on statuses at a given time, using the techniques of user's pattern analysis, anomaly detection, textual topic detection and social network analysis for Counter-Terrorism efforts.
Then, Resig et al. study on IM communication as social networks trying to measure the relation between members by status log of an IM user, state and time at which member switched into that state using IMSCAN framework [32].
They compare the social network obtained using the relationship measures to the social networks formed in LiveJournal. They study IM communities as social networks. So they use the status log as the basis of a measure of the degree to which any two AOL IM users are related. The status log of an IM user is a list of pairs of the form (time, state), where state is an element of a small set, such as *online, offline, busy, away*, and time is the time at which the member switched into that state.
They show that, in spite of their simplicity, status logs contain a great deal of structure. Since any pair of IM users can instant message each other only if they are both online at the same time, it seems reasonable to guess that any two IM users that are frequently online at the same time may in fact be frequently instant messaging each other.
For a chosen population of IM users, they compare the social networks obtained using their relationship measures to the social network formed in LiveJournal by the same population. LiveJournal is a blogging community that allows users to explicitly name other LiveJournal users as associates.
The network obtained by these association lists thus acts as a control of sorts for validating our IM-based association measure. They describe two experiments: in the first, for each user in the LiveJournal graph, they compare the number of seconds that user was online on AOL IM to the out-degree of the node in the LiveJournal graph corresponding to that user. In the second, for each pair of users in the LiveJournal graph (not necessarily linked), they measure the degree to which the pair's IM online is synchronized to whether the pair is linked in the

LiveJornal graph. They also study techniques to recovering associations between users from their observed behavior by first clustering them based on the status logs.

Kucukyilmaz et al. investigate the possibility of predicting various author-specific and message-specific attributes in chat environments using machine learning techniques [23].
A term-based approach is used to investigate the user and message attributes in the context of vocabulary use while a style-based approach is used to examine the chat messages according to the variations in the authors' writing styles.

Orebaugh and Allnutt in 2009 determined if an author of an IM conversation could be identified based on his or her sentence structure and use of special characters, emoticons, and abbreviations for forensic analysis [30].
The experiments also determined which features were strongest at identifying authors for the purpose of computer forensics analysis. On 69 stylometric features (17 special characters, 16 emoticons, 35 abbreviations) the strongest identifying attributes according to the information gain and chi-squared techniques are 'U', three dots and hypen.
For possible variation each profile need a degree of confidence. The research use the Weka data mining tool for classification. The classification methods used are j48 decision tree, IBK nearest neighbor and Naive Bayes classifiers.
They show the accuracy, error, true positive and false positive rate for each classifier applied to the data set. They show also the classification accuracy results for attributes category.

Based on previous work in 2006 [29], in which the focus is to masquerade attack from another user, every conversation is logged in a text file that was easy to parse and analyze.
The framework, first apply pattern analysis to the collected conversation for extract a sort of user profile based on IM characteristics. Next, the framework apply an anomaly *intrusion detection system*. Using the profiles collected, users, whose actions deviate from profile, were flagged.
They analyse histograms of single user conversation, multiple user conversation,

showing outliers and features frequency distributions of uppercase characters, lowercase characters, numbers and special characters.
In the first test they compute mean and standard deviation to determining the data in which the user maintain a relative consistency for a particular feature.
In the second test they compute mean and standard deviation for the character frequency diagrams for each feature, for each user to assist in distinguishing among authors. Standard deviation provides a stylometric measure per user for profiling building, for example the larger the standard deviation, the more variability to identify differences between users. It is used also to determine outliers for each user by using the empirical rule. Thus, they compute standard Z score for each user for each characters and it measures the number of standard deviation above or below the mean. Then, they use the measure of certain threshold for the standard score to determine outliers.
In the third experiment they compare new unclassified conversation to the learned user profiles (instance-based learning): it use a distance function to determine which user in training set is closest to the unknown one (KNN). Then a degree of confidence is generated for each attribute on the value of distance.

Ali et al. [7] investigated on identification of Chat Bots - a computer application designed to simulate a conversation with a human user - by their style, showing average accuracy for each feature over all classifiers they considered, to overcome the danger of online criminal activities, analyzing conversion log files. The collected data comes from chat logs between different Chat Bots and between Chat Bots and human users. They use, among other features, #characters, #words and word length by mean of each feature.

Abbasi et al. [3, 4], Iqbal et al. [19], Stamatatos [36] and Zheng et al. [43] studies are the most standard and cited works on authorship attribution using stylometric analysis. In their works investigate in authorship identification and attribution of written documents, drafting a taxonomy for online stylometric analysis.

More specifically, Zheng et al. develop a framework for authorship identification of online messages to address the identity-tracing problem. Four type of writing

style features (lexical, syntactic, structural and content-specific) are extracted (see Figure 2.1) and inductive learning algorithms are used to build feature-based classification models to identify authorship of online messages.
They compared the discriminating power of the four types of features and of three classification techniques: decision trees, back-propagation neural networks, and support vector machines.
They use a small set of authors and validate the features accuracy on 30 random messages. Function words, part of speech and punctuation usage are considered as syntactic features. Paragraph length, use of indentation, use of signature are instead classify as structural features. Content-related words or phrases as content-specific features.
Usually, these features can express personal interest in a specific domain. In the experiments they use an incremental features set, i.e. they first consider lexical features set, then they add syntactic, then content-specific ones; this in order of evolutionary sequence of style features.
To evaluate the prediction, they used the accuracy measure, which has been commonly adopted in data mining and authorship analysis. Accuracy indicates the overall prediction of a particular classifier, defined as accuracy of messages whose author was correctly identified per total number of messages.

Abbasi et al. in 2008 incorporate a rich set of features (see Figure 2.2), including lexical, syntactic, structural, content-specific and idiosyncratic attributes to develop writeprints technique for identification and similarity detection of anonymous identities.
The concept of writeprint, an analogy of a fingerprint in physical forensic analysis, is to capture the writing style of a person from his/her written text. They propose the use of stylometric analysis to help identify online traders based on the writing style traces inherent in their posted feedback comments.
Experiments conducted to assess the scalability (number of traders) and robustness (against intentional obfuscation) of the proposed approach found it to significantly outperform benchmark stylometric techniques. Scalability refers to the impact of the number of author classes on classification performance. It is also important to assess robustness of stylometric approaches against intentional stylistic alter-

Features	Description
Lexical features	
Character-based features	
1. Total number of characters(C)	
2. Total number of alphabetic characters/C	
3 Total number of upper-case characters/C	
4. Total number of digit characters/C	
5. Total number of white-space characters/C	
6. Total number of tab spaces/C	
7–32. Frequency of letters (26 features)	A–Z
33–53. Frequency of special characters (21 features)	~, @, #, $, %, ^, &, *, -, _, =, +, >, <, [,], {, }, /, \, |
Word-based features	
54. Total number of words (M)	
55. Total number of short words (less than four characters)/M	e.g., and, or
56. Total number of characters in words/C	
57. Average word length	
58. Average sentence length in terms of character	
59. Average sentence length in terms of word	
60. Total different words/M	
61. Hapax legomena*	Frequency of once-occurring words
62. Hapax dislegomena*	Frequency of twice-occurring words
63. Yule's K measure*	A vocabulary richness measure defined by Yule
64. Simpson's D measure*	A vocabulary richness measure defined by Simpson
65. Sichel's S measure*	A vocabulary richness measure defined by Sichele
66. Brunet's W measure*	A vocabulary richness measure defined by Brune
67. Honore's R measure*	A vocabulary richness measure defined by Honore
68–87. Word length frequency distribution /M (20 features)	Frequency of words in different length
Syntactic Features	
88–95 Frequency of punctuations (8 features)	",", ".", "?", "!", ":", ";", "'", """
96–245 Frequency of function words (303 features)	The whole list of function words is in the appendix.
Structural Features	
246. Total number of lines	
247. Total number of sentences	
248. Total number of paragraphs	
249. Number of sentences per paragraph	
250. Number of characters per paragraph	
251. Number of words per paragraph	
252. Has a greeting	
253. Has separators between paragraphs	
254. Has quoted content	Cite original message as part of replying message
255. Position of quoted content	Quoted content is below or above the replying body
256. Indentation of paragraph	Has indentation before each paragraph
257. Use e-mail as signature	
258. Use telephone as signature	
259. Use url as signature	
Content-specific Features	
260–270. Frequency of content specific keywords (11 features)	"deal", "obo", "sale", "wtb", "thx", "paypal", "check", "windows", "software", "offer", "Microsoft"

Figure 2.1: Adopted feature in the framework developed by Zheng.

ation and copycatting/message forging. Fraudulent traders may attempt to avoid detection by altering their style or copying other traders' style (referred to as copy-

catting or forging).
Forging/copycatting entails intentionally mimicking other community members' styles or user names. This behavior is fairly common in certain computer-mediated communication (CMC) modes, such as Usenet forums. Mimicking other members' styles by either directly copying their text or attempting to copy their stylistic tendencies is an important and plausible form of deception that must be considered when evaluating stylometric methods in online settings.
They developed the writeprint technique that uses Karhunen–Loeve transforms and a novel pattern disruption mechanism to help detect stylistic similarity between traders based on feedback comments. Experiments in comparison with existing stylometric techniques demonstrated the scalability and robustness of the proposed features and technique for differentiating trader identities in online markets.
In the same year Abbasi and Chen worked in a similar topic including email, IM, feedback comments and program code.

Group	Category	Quantity	Description/examples
Lexical	Word level	5	Total words, percent characters per word
	Character level	5	Total characters, percent characters per message
	Character n-grams	< 18,278	Count of letter n-grams (e.g., a, at, ath)
	Digit n-grams	< 1,110	Count of digit n-grams (e.g., 1, 12, 123)
	Word-length distribution	20	Frequency distribution of 1–20 letter words
	Vocabulary richness	8	Richness (e.g., hapax legomena, Yule's K)
	Special characters	21	Occurrences of special characters (e.g., @#$%^&*+=)
Syntactic	Function words	300	Frequency of function words (e.g., of, for, to)
	Punctuation	8	Occurrence of punctuation marks (e.g., !;:,.?)
	Part-of-speech tag n-grams	Varies	Part-of-speech tag n-grams (e.g., NNP, NNP JJ)
Structural	Message level	6	For example, has greeting, has URL, requoted content
	Paragraph level	8	For example, number of paragraphs, sentences per paragraph
	Technical structure	50	For example, file extensions, fonts, use of images
Content specific	Word n-grams	Varies	Bag-of-word n-grams (e.g., "seller", "bad sale")
Idiosyncratic	Misspelled words	< 5,513	Common misspellings (e.g., "beleive", "thougth")

Figure 2.2: Extracted feature set by Abbasi.

Stamatatos analysed previous works on authorship attribution that have proposed taxonomies of features to quantify the writing style under different labels and criteria (see Figure 2.3).
He reviewed the text representation features for stylistic purpose focusing on the computational requirements for measuring them. A survey of recent advances of

the automated approaches to attributing authorship is presented examining their characteristics for both text representation and text classification.
The focus of this survey is on computational requirements and settings rather than linguistic or literary issues. He also discusses evaluation methodologies and criteria for authorship attribution studies and list open questions that will attract future work in this area.

Features		Required tools and resources
Lexical	Token-based (word length, sentence length, etc.)	Tokenizer, [Sentence splitter]
	Vocabulary richness	Tokenizer
	Word frequencies	Tokenizer, [Stemmer, Lemmatizer]
	Word *n*-grams	Tokenizer
	Errors	Tokenizer, Orthographic spell checker
Character	Character types (letters, digits, etc.)	Character dictionary
	Character *n*-grams (fixed length)	–
	Character *n*-grams (variable length)	Feature selector
	Compression methods	Text compression tool
Syntactic	Part-of-speech (POS)	Tokenizer, Sentence splitter, POS tagger
	Chunks	Tokenizer, Sentence splitter, [POS tagger], Text chunker
	Sentence and phrase structure	Tokenizer, Sentence splitter, POS tagger, Text chunker, Partial parser
	Rewrite rules frequencies	Tokenizer, Sentence splitter, POS tagger, Text chunker, Full parser
	Errors	Tokenizer, Sentence splitter, Syntactic spell checker
Semantic	Synonyms	Tokenizer, [POS tagger], Thesaurus
	Semantic dependencies	Tokenizer, Sentence splitter, POS tagger, Text Chunker, Partial parser, Semantic parser
	Functional	Tokenizer, Sentence splitter, POS tagger, Specialized dictionaries
Application-specific	Structural	HTML parser, Specialized parsers
	Content-specific	Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries
	Language-specific	Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries

Figure 2.3: Stamatatoes' types of stylometric features together with computational tools and resources required for their measurement.

Finally, Iqbal et al. present a unified data mining approach to address the challenges of authorship attribution in anonymous online textual communication (email, blog, IM) for the purpose of cybercrime investigation.
They study three problem of authorship analysis problem: authorship identification with large (text) and small (messages) training set and authorship characterization including lexical, syntactic, structural, domain-specific, gender-preferential features.
The contributions are summarized as frequent-pattern-based writeprint, capturing stylistic variation (a person may have multiple writing styles depending on the recipients and the context of a text message). The problem of the size of the training set has been studied, with different classifiers and 20 subjects.

Based on the review of previous studies and analysis, five types of features into the feature set are integrated : lexical, syntactic, content-specific, and structural features, idiosyncratic features.
Table 2.4 shows features referred to previous AA chats works with respect to the five type of writing style.

Group	Description	Examples	References
Lexical	Word level	Total number of words (=M), # short words/M, # chars in words/C, # different words, chars per word, freq. of stop words	[4, 7, 19, 23, 30, 36, 43]
	Character level	Total number of characters (chars) (=C), # uppercase chars/C, # lowercase chars/C, # digit chars/C, freq. of letters, freq. of special chars	[4, 23, 30, 36, 43]
	Character—Digit n-grams	Count of letter—digit n-gram (a, at, ath, 1 , 12 , 123)	[4, 7, 23, 36, 43]
	Word-length distribution	Histograms, average word length	[4, 7, 19, 23, 30, 36, 43]
	Vocabulary richness	Hapax legomena, dislegomena	[4, 7, 19, 36, 43]
Syntactic	Function words	Frequency of function words (of, for, to)	[4, 7, 19, 30, 36, 43]
	Punctuation	Occurrence of punctuation marks (!, ?, :), multiple !—?	[4, 19, 23, 30, 36, 43]
	Emoticons—Acronym	:-), L8R, Msg, :(, LOL	[23, 30, 36]
Structural	Message level	Has greetings, farewell, signature	[4, 19, 30, 36, 43]
Content-specific	Word n-grams	Bags of word, agreement (ok, yeah, wow), discourse markers—onomatopee (ohh), # stop words, # abbreviations , gender—age-based words, slang words	[4, 7, 19, 23, 30, 36, 43]
Idiosyncratic	Misspelled word	Belveier instead of believer	[4, 19, 30, 36]

Table 2.4: Synopsis of the state-of-the-art features for AA on chats. "#" stands for "number of".

Chapter 3

Mathematical Background

Contents

Stylometric features are typically extracted from the data and fed into discriminative classifiers, where each author is a class.

The majority of typical research works in stylometry, apply an instance-based approach to achieve AA. A typical architecture of such an instance-based approach is shown in Figure 3.1.

In detail, each text sample of the training corpus is represented by a vector of attributes (x) and a classification algorithm is trained using the set of instances of known authorship (training set) in order to develop an attribution model. Then, this model will be able to estimate the true author of an unseen text.

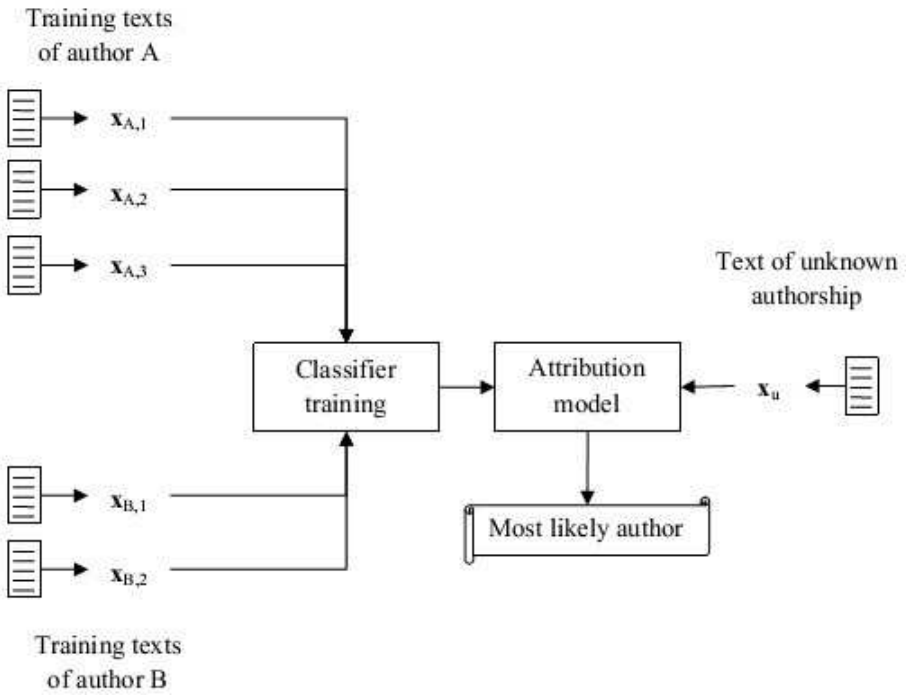


Figure 3.1: Typical instant-based approach

Our approach does not exploit any classifiers, but relies on a feature extraction and selection phase, that gives a ID signature for each individual of the gallery set. Once a test sample is given, the test ID signature is extracted and compared with all the individuals' ID signatures in the gallery. The comparison consists in a minimization step, in which the distances among signatures are ranked, from the nearest one to the farthest. If the right correspondence is found in the early position of the rank, the features are expressive, and we achieved our task. The architecture of this profile-based approach is shown in Figure 3.4.

As the number of features increases, so does the amount of information. While having a large number of features may seem preferable, it is possible that the performance gets worse with more than with fewer features.
With enough samples we are essentially assured of convergence. On the other hand, the number of samples needed may be very large indeed, much grater than would be required. Little or nothing in the way of data reduction is provided,

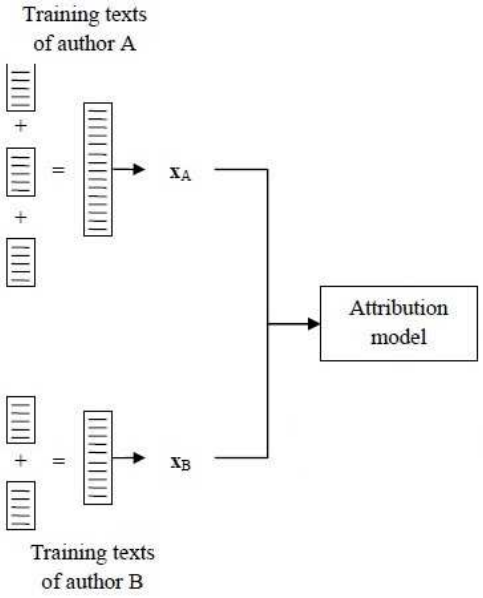


Figure 3.2: Training phase

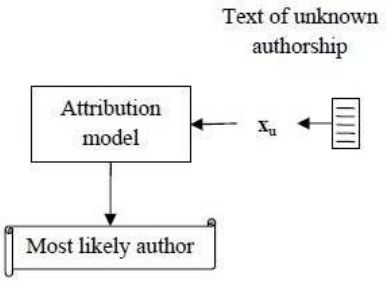


Figure 3.3: Testing phase

Figure 3.4: Typical profile-based approach

which leads to severe requirements for computation time and storage. Moreover, the demand for large number of samples grows exponentially with the dimensionality of the feature space.
This is often term "curse of dimensionality", a term coined by Richard Bellman and restricts the practical application of the procedure. The fundamental reason for the curse of dimensionality is that high-dimensional functions have the potential to be much more complicated than low-dimensional ones, and that those complications are harder to discern.
The only way to beat the curse is to incorporate knowledge about the that is correct.

Fortunately, this curse of dimensionality can be ameliorated by proper selection and reduction of features. We present a method for determining which features are the most important or salient in next sections. Knowing the salient features allows to reduce the dimensionality of the input data by eliminating poor features.

Feature extraction and selection are two approaches for dimension reduction.

Feature extraction — Combining attributes into a new reduced set of features.
Feature selection — Selecting the most relevant attributes.

3.1 Feature extraction

In pattern recognition and in image processing, feature extraction is a special form of dimensionality reduction.

Selfridge [25] defines pattern recognition solely in terms of "the extraction of the significant features from a background of irrelevant detail."
It is emphasized, and this is an important point, that the significance is a function of both context and the experience of the pattern recognizer. On the subject of feature extraction, Nilsson comments that:

1. No general theory exists to allow us to choose what features are relevant for a particular problem
2. Design of feature extractors is empirical and uses many ad hoc strategies
3. We can get some guidance from biological prototypes.

When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant (much data, but not much information) then the input data will be transformed into a reduced representation set of features (also named features vector).
Transforming the input data into the set of features is called feature extraction. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input.

Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. The feature extraction process results in a much smaller and richer set

of attributes.
Feature extraction can be used to extract the themes of a document collection, where documents are represented by a set of key words and their frequencies. Each feature is represented by a combination of keywords. The documents in the collection can then be expressed in terms of the discovered features.

The most important criterion for selecting features in authorship attribution tasks is their frequency. In general, the more frequent a feature, the more stylistic variation it captures. Houvardas and Stamatatos (2006) proposed an approach for extracting character n-grams of variable length using frequency information only. The comparison of this method with information gain, a well-known feature selection algorithm examining the discriminatory power of features individually (Forman, 2003), showed that the frequency-based feature set was more accurate for feature sets comprising up to 4.000 features.

3.2 Feature selection

Feature selection (also known as subset selection) is a process commonly used in machine learning, wherein a subset of the features available from the data are selected for application of a learning algorithm.

The problem of feature selection is defined as follows: given a set of candidate features, select a subset that performs the best under some classification system. The term feature selection is taken to refer to algorithms that output a subset of the input feature set.

This procedure can reduce not only the cost of recognition by reducing the number of features that need to be collected, but in some cases it can also provide a better classification accuracy due to finite sample size effects.
In authorship studies too, one may apply different feature selection techniques to determine a subset of stylometric features that can discriminate the authors.
The best subset contains the least number of dimensions that most contribute to accuracy. This is an important stage and is one of two ways of avoiding the curse

of dimensionality (the other is feature extraction).

Simple feature selection algorithms are ad hoc, but there are also more methodical approaches. From a theoretical perspective, it can be shown that optimal feature selection for supervised learning problems requires an exhaustive search of all possible subsets of features of the chosen cardinality. If large number of features are available, this is impractical. For practical supervised learning algorithms, the search is for a satisfactory set of features instead of an optimal set.

Feature selection algorithms typically fall into two categories: *feature ranking* and *subset selection*.
Feature ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score. Subset selection searches the set of possible features for the optimal subset.

In statistics, the most popular form of feature selection is stepwise regression. It is a greedy algorithm that adds the best feature (or deletes the worst feature) at each round. The main control issue is deciding when to stop the algorithm.
In machine learning, this is typically done by cross-validation. In statistics, some criteria are optimized. This leads to the inherent problem of nesting. More robust methods have been explored, such as branch and bound and piecewise linear network.
For a summary of feature selection method see Figure 3.5.

3.2.1 Feature selection as heuristic search

One can view the task of feature selection as a search problem, with each state in the search space specifying a subset of the possible features. As Figure 3.6 depicts, one can impose a partial ordering on this space, with each child having exactly one more feature than its parents. The structure of this space suggests that any feature selection method must take a stance on four basic issues that determine the nature of the heuristic search process.

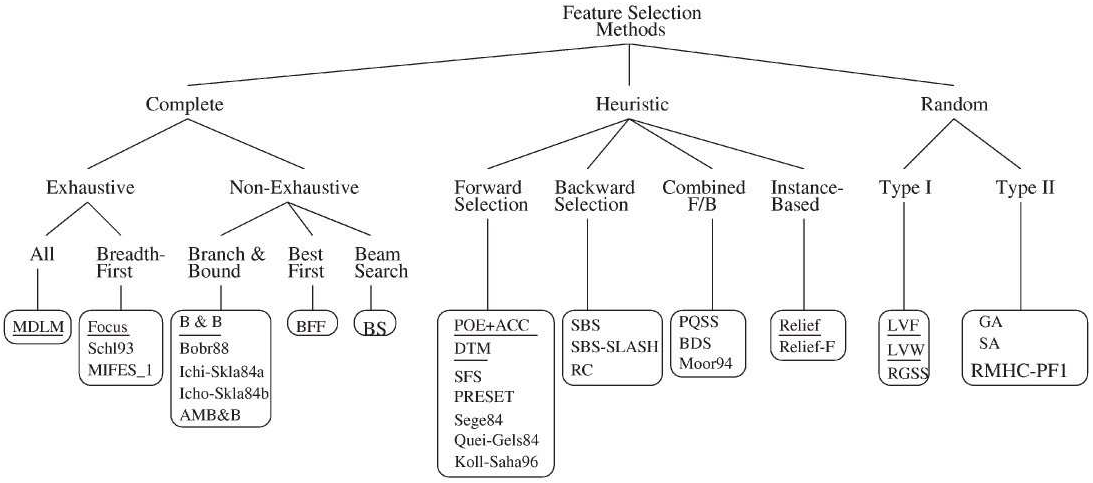


Figure 3.5: Summary of feature selection methods. Dash and Liu (1997)

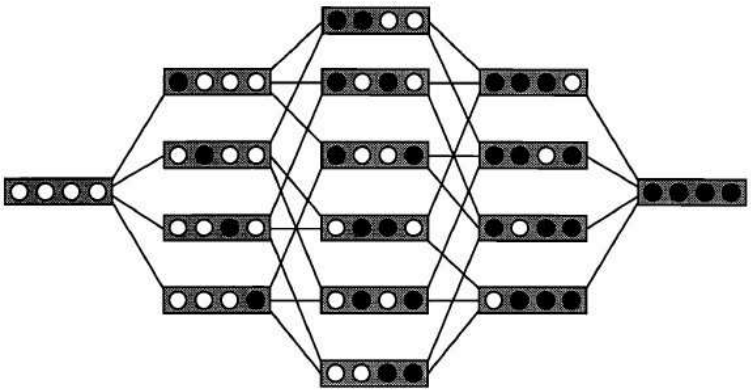


Figure 3.6: Each state in the space of feature subsets specifies the attributes to use during induction. Note that the states in the space (in this case involving four features) are partially ordered, with each of a state's children (to the right) including one more attribute (dark circles) than its parents.

First, one must determine the starting point in the space, which in turn determines the direction of search. For instance, one might start with no features and successively add attributes, or one might start with all attributes and successively remove them. The former approach is sometimes called forward selection, whereas the latter is known as backward elimination.

Forward selection (sequential forward selection) starts with no features and, at each step, adds the feature that decreases the error the most until any further addition does not decrease the error significantly. Such approach is used in this research.
Backward selection starts with all the features and, at each step, removes the one that decreases the error the most until any further removal increases the error significantly. These approaches consider only one attribute at a time. One might also select an initial state somewhere in the middle and move outward from this point.

A second decision involves the organization of the search. A more realistic approach relies on a greedy method to traverse the space. At each point in the search, one considers local changes to the current set of attributes, selects one, and then iterates, never reconsidering the choice. A related approach, known as stepwise selection or elimination, considers both adding and removing features at each decision point, which lets one retract an earlier decision without keeping explicit track of the search path. Within these options, one can consider all states generated by the operators and then select the best, or one can simply choose the first state that improves accuracy over the current set.

A third issue concerns the strategy used to evaluate alternative subsets of attributes. One broad class of strategies considers attributes independently of the induction algorithm that will use them, relying on general characteristics of the training set to select some features and exclude others. John, Kohavi, and Pfleger (1994) call these filter methods, because they filter out irrelevant attributes before the induction process occurs. They contrast this approach with wrapper methods, which generate a set of candidate features, run the induction algorithm on the training data, and use the accuracy of the resulting description to evaluate the feature set. Within this approach, one must still pick some estimate for accuracy, but this choice seems less central than settling on a filter or wrapper scheme.

Finally, one must decide on some criterion for halting search through the space of feature subsets. Within the wrapper framework, one might stop adding or remov-

ing attributes when none of the alternatives improves the estimate of classification accuracy, one might continue to revise the feature set as long as accuracy does not degrade, or one might continue generating candidate sets until reaching the other end of the search space and then select the best. Within the filter framework, one criterion for halting notes when each combination of values for the selected attributes maps onto a single class value. Another alternative simply orders the features according to some relevancy score, then uses a system parameter to determine the break point.

The feature selection approach used in this research is explained in the following.

3.2.2 Sequential forward selection

Sequential forward selection (SFS) is the simplest greedy search algorithm. The estimate of the quality of the candidate subsets usually depends on the training/testing split of the data.

Definition 3.1 (SFS)**.** *Let $X = x_1, \ldots, x_n$ be the original set of features and $J(S)$ be a measure of quality of a subset $S \subseteq X$. In particular $J(S)$.*
Starting with an empty subset, S, one feature is added at each step. To choose this feature, all possible subsets of $S \cup x_i$ are evaluated, where x_i is a feature from X which is not in S. The best feature to add is taken to be

$$x^* = argmax_{x_i \in X \setminus S} \quad J(S \cup \{x_i\}).$$

Starting from the empty set, sequentially add the feature x^* that results in the highest objective function $J(S \cup \{x_i\})$ when combined with the features $S \cup x_i$ that have already been selected.

The algorithm can be summarized in four step:
1. Start with the empty set $S_0 = \emptyset$
2. Select the next best feature $x^* = argmax_{x \notin S_k} \quad S_k + x^*$.
3. Update $S_{k+1} = S_k + x^*; \quad k = k + 1$
4. Go to 2

SFS performs best when the optimal subset has a small number of features.

As an example, the state space for 4 features is shown Figure 3.7. Notice that the number of states is larger in the middle of the search tree.

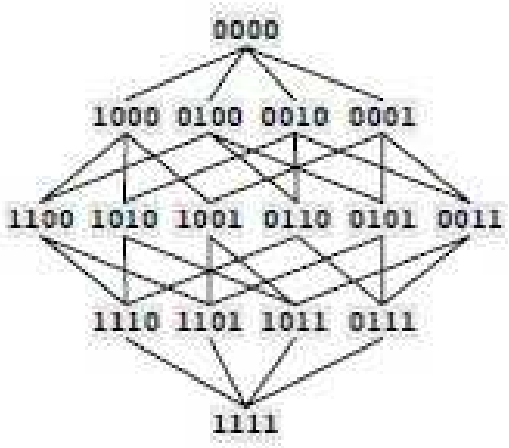


Figure 3.7: Search space

The main disadvantage of SFS is that it is unable to remove features that become obsolete after the addition of other features.

3.2.3 A stability Index

It has been well documented that a subset of features may work better than the entire set.
Selecting a suitable subset of features is not only computationally desirable but can also lead to better classification accuracy. The quality of a feature subset is measured by an estimate of the classification accuracy of a chosen classifier trained on the candidate subset.

Therefore, when two feature subsets are compared, the decision as to which one should be preferred involves uncertainty. This is particularly important in the sequential feature selection methods which augment or reduce the selected subset

at each step. A flip in the decision at an earlier step may lead to a completely different selection path, and result in a very different subset of features being selected. Stability of the selected features is an important aspect when the task is knowledge discovery, not merely returning an accurate classifier.

As different sequences of features may be returned from repeated runs of SFS, a substantial discrepancy between such sequences can signal a problem with the selection. A stability index is suggested based on cardinality of the intersection and a correction for chance. The experimental results that Kuncheva achieved, indicate that the index can be useful for selecting the final feature subset. If stability is high, then we should return a subset of features based on their total rank across the SFS runs. If stability is low, then it is better to return the feature subset which gave the minimum error across all SFS runs.

Consider the sequential forward selection (SFS) procedure. The problem is that we do not have the exact value of $J(S \cup \{x_i\})$ but only an approximation thereof evaluated on a part of the training data.
Thus the choice of x^* depends on the accuracy of this estimate. If a large training set is available or if one can afford a large number of data shuffling runs, so that the variance of $J(S \cup \{x_i\})$ is small, the estimate will be reliable enough and the choice of x^* will be unequivocal.

However, when this is not possible, we have to account for the variability of $J(S \cup \{x_i\})$ in other ways. If SFS is run to the end, the result is a sequence of features entering the subset. If a subset of d features is required, the first d features of the sequence will be returned. Suppose that we carry out K runs of SFS and record the sequences $S_1, S_2, \ldots, S_K$. The question is how similar these sequences are and whether this similarity can help us choose the final subset to be returned to the user.

Let $S_1, S_2, \ldots, S_K$ be the sequences of features obtained from K runs of SFS on a given dataset.

Definition 3.2 (Stability index for K sequences)**.** *The Stability Index for a set of sequences of features, $\mathcal{A} = S_1, S_2, \ldots S_K$, for a given set size, k, is the average of all pairwise consistency indices*

$$\mathcal{I}_S(\mathcal{A}(k)) = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} I_c(S_i(k), S_j(k)).$$

3.2.4 Choosing final sequence of features

As different sequences of features may be returned from repeated runs of SFS, in order to cut a subset of features to return to the user, a final sequence S^* has to be chosen. Given $\mathcal{A}$, there are various intuitive options for choosing S^* and the number of features. The following two options comply with the current practices.

- Rank the features in each sequence S_i so that the best feature (the one starting the sequence) is assigned rank 1, the second is assigned rank 2, etc. Sum up the K ranks for each feature. Order the features in ascending order by the total ranks to get a final sequence S^*_{rank}.
- Find the sequence with the minimum local error.

$$S^*_{min} = argmin_i\{min_k J(S_i(k))\}.$$

Having a set of sequences instead of a single one opens up a multitude of choices with respect to selecting the number of features d. Some of the possible way to pick d are

(a.) Using the final sequence, pick the number d such that $d = argmin_k J(S^*(k))$.

(b.) Using the final sequence, fit a polynomial to $J(S^*(k))$ as a function of k. Find the minimum analytically using the coefficients. The integer value of k closest to the minimum is retrieved as d. The benefit of this calculation is that the criterion curve will be smoothed. Fluctuations of $J(S^*(k))$ are expected to occur due to estimation errors. Such fluctuations will represent noise in the selection process and so should be eliminated.

(c.) Apply (a) on the mean error across the K sequences.

(d.) Apply (b) on the mean error across the K sequences.

(e.) Find the suggested number of features for each S_i by either (a) or (b). Let d_i be this number for sequence $S_i, i = 1, \ldots, K$. Compare all d_i and derive a final d based on median, mode or mean.

(f.) Find a set of "consistent values" d, for which the sequences agree, i.e.,

$$D_{cons} = d|1 \leq d \leq n, \mathcal{I}_S(\mathcal{A}(d)) > \Theta,$$

where Θ is a predefined threshold on $\mathcal{I}_S$. Choose $d$to be the one with the smallest $J(S^*(d))$ within D_cons.

If the final sequence is S^*_{min}, then picking d is straightforward because the errors for $k = 1, 2, \ldots, n$ are available from the training run producing S^*_{min}. If S^*_{rank} is chosen, another evaluation run has to be carried out in order to find the validation error of each subset S^*_{rank}.

Intuitively, if stability is high, S^*_{rank} would be better because it will smooth out the small discrepancies between the selected subsets. If stability is low, perhaps there have been runs which have discovered by chance irregular troughs of the error criterion J, accounting for a set of dependent and useful features. In that case it may be better to use S^*_{min} and return the best feature subset across all individual runs.

Chapter 4

Method

Contents

A set of novel stylometric features that take into account the conversational nature of chat interactions is proposed. Some of them, lexical and syntactic features, fit in the taxonomy proposed in the literature, but others require to define a new group of features, here called *conversational* features.

Stylometric features are typically extracted from the data and use discriminative classifiers to identify the author (each author corresponds to a class). The extraction process is always applied to the entire conversation and the individual turns, while being the basic blocks of the conversation, are never used as analysis unit.

In this work we try to improve the effectiveness of the previous AA approaches. We introduce two novelties: we adopted features inspired by turn-taking in the conversation, and we try to extract the features from individual turns rather than from entire conversations.

The reason is that they are based on *turn-taking*, probably the most salient aspect of spoken conversations that applies to chat interactions as well. In conversations, *turns* are intervals of time during which only one person talks. In chat interactions, a turn is a block of text written by one participant during an interval of time in which none of the other participants writes anything. Like in the case of automatic analysis of spoken conversations, the AA features are extracted from individual turns and not from the entire conversation.

In next section we explain how Skype client works and how our conversation are stored in a log file. Thus, we show our feature extraction method.

4.1 Skype

We used the Skype client to collect data from IM conversation.

Voiceover IPs are famous for their ability to connect voice calls over Internet lines. Skype goes a step further by enabling a fully functional chat feature.
There is a chat saving feature so you'll always know who said what. Chat archives are stored on your computer, not on the network. The only people who will ever be able to see what you said during a Skype chat, are those that were in the chat group with you.

Skype is a proprietary voice-over-Internet Protocol service and software application. The service allows users to communicate with peers by voice, video, and instant messaging over the Internet. Phone calls may be placed to recipients on the traditional telephone networks. Unlike most VoIP services, Skype is a hybrid peer-to-peer and client–server system, and makes use of background processing on computers running Skype software; the original name proposed – Sky peer-to-peer – reflects this.
Registered users of Skype are identified by a unique Skype Name, and may be listed in the Skype directory. Skype allows these registered users to communicate through both instant messaging and voice chat.

Voice chat allows telephone calls between pairs of users and conference calling, and uses a proprietary audio codec. Skype's text chat client allows group chats, emoticons, storing chat history and editing of previous messages. The usual features familiar to instant messaging users - user profiles, online status indicators, and so on - are also included.
Skype interface is shown in Figure 4.1.

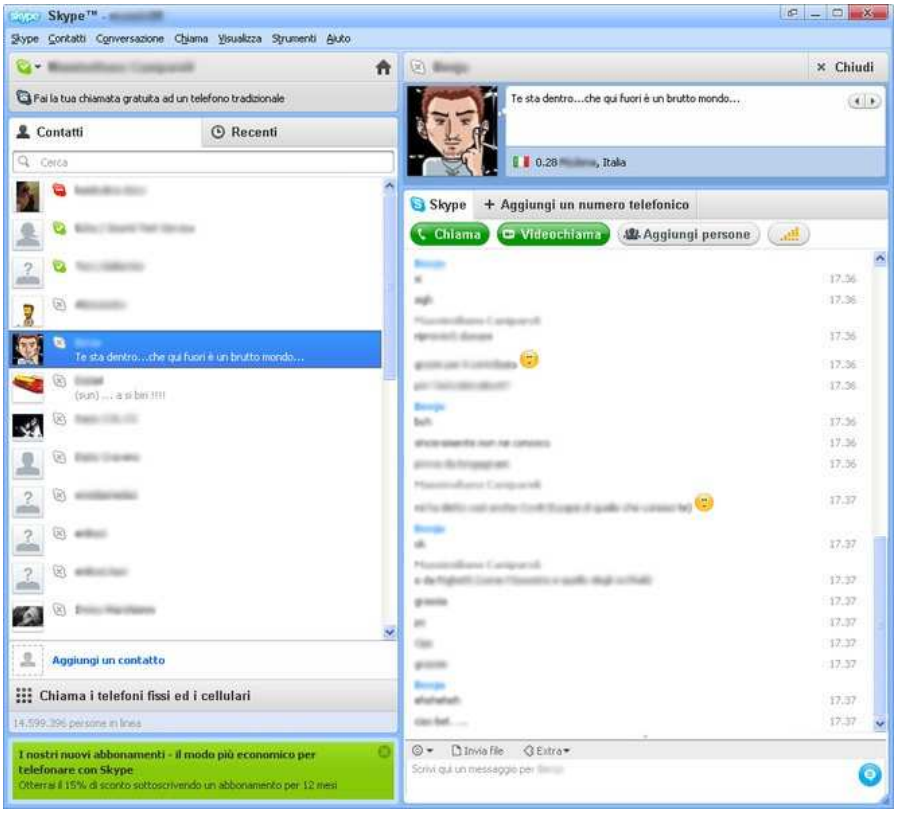


Figure 4.1: Skype interface

4.1.1 Where is my Skype chat conversation history stored?

Skype software uses a number of files to store data. These files relate mainly to historical information, call histories, file transfers, messaging sessions, etc. They also cache user profiles. The interpretation of these log files can yield a significant

amount of information about communications that have taken place through the software.

The default settings of the Skype application are set to store chat conversation history forever. When you have chat conversations on Skype, these are stored on the computer that you are "chatting" on in a folder that has the same name as your "Skype Name", the location for which varies depending on the operating system being used.

Under Skype for Windows, user data files are stored in the user's Application Data folder under the Skype subfolder; this is then subdivided based on Skype user, allowing multiple Skype users to operate under the same Windows account. So, your Skype chat history is stored in a file called *main.db* which resides at the following location on a Windows 7 Operating system:
C:\Users\[User Name]\AppData\Roaming\Skype\[Skype Name]\.

In the following are listed the information available for extraction from Skype logs about messages. Note that the sequence number allows the order of events to be determined, without relying on the resolution of the timestamp.

Messages (e.g. msg256.dbb or chatmsg256.dbb)

- Sequence Number
- Message content
- Chat ID (groups messages within a chat session)
- Timestamp
- User name (sender)
- Display name (sender)

Files are stored with a .dbb extension with the filename consisting of a string describing the contents followed by a number which indicates the record length (e.g. call256.dbb, chatmsg512.dbb etc). The minimum record length observed is 256 bytes, with files seen up to 16384 bytes. Items are stored in the smallest length format possible with blank padding to fill any space remaining in the

record. Therefore it is quite common to have multiple files with the same prefix and different record lengths.

The message file contains the following data items:

Data item indicator	**Data item**	**Format**
0xFC 0x03	Message content	Null-terminated string
0xE0 0x03	Message ID	Null-terminated string
0xE5 0x03	Time stamp	As described above
0xE8 0x03	User name (sender)	Null-terminated string
0xEC 0x03	Display name (sender)	Null-terminated string

Table 4.1: Message data items

The Message ID is a string which uniquely identifies a chat session. Consequently the thread of messages can be assembled from grouping those items with identical message IDs.

In the following, we discuss our feature extraction approach, taking into account the new conversational features. We explain their means and how we compute them from the raw features, extracted from chat conversations.

4.2 Data preparation

A feature set is composed by writing-style features predefined by us. As an important component of our research, the feature set may significantly affect the performance of authorship identification.

For the special characteristics of chat messages discussed earlier, new feature selection heuristics are necessary. Based on the review of previous studies and analysis, we have seen that five types of features into the features set are integrated: lexical, syntactic, structural features, content-specific, idiosyncratic features.

Privacy and ethical issues limit the use of the features describe in previous taxonomy. Only those features that do not involve the content of the conversation can

be used, namely number of words, characters, punctuation marks and emoticons. In standard AA approaches, these features are counted over entire conversations, obtaining a single quantity. In our case, we consider the turn as a basic analysis unit, so we extract such features for each turn.
We will use only a small part of static features, related to lexical and syntactic features.

4.2.1 Feature extraction

Each subject who participated to our experiments, collected his/her message history, storing each conversation in a sperate .txt file in the following format.

> [24/04/2012 13:37:59] Mario Rossi: Ciaoooo, come va??
> [24/04/2012 13:38:09] Chiara Bianchi: Hey, tutto alla grande!

Our dataset includes $N = 77$ subjects, each involved in a dyadic chat conversation with an interlocutor. The feature extraction process is applied to T consecutive turns that a subject produces during the conversation. Since, in our case, we consider the turn as a basic analysis unit, so we extract such features for each turn, obtaining T numbers.

We firstly apply ***feature extraction*** that takes information we are interested in, from each conversation and provide a structure aims to achieve our dataset, i.e raw features. No information about real conversations will be stored.
We consider that new conversation can start after 30 minutes of conversation inactivity.

Raw features extracted from log files are :

1 - number of words in this turn

2 - number of emoticons in this turn

3 - number of exclamation points

4 - number of characters

5 - response time

6 - number of pressed return

7 - imitation rate (length(msg_A)/length(msg_B)

8 - answer time, if a question mark was detected before

9 - number of question marks

10 - number of '...'

11 - number of only 'UpperCase Letters'

Then, from these raw features we build our complete features set, that includes also conversational features.

1 - number of return per turn (return chars) $\star$

2 - number of words per turn

3 - number of emoticons per turn

4 - number of emoticons per word

5 - number of emoticons per characters

6 - number of exclamation mark per turn

7 - number of question mark per turn

8 - number of characters per turn

9 - number of words/characters rate per turn (average word length)

10 - rate of mimicry per turn $\star$

11 - number of three points per turn

12 - number of Upper case letters per turn

13 - number of uppercase letters per used words

14 - resptime (turn duration) $\star$

15 - writing speed (chars per second) $\star$

16 - number of word per second $\star$

Those features flagged with ⋆ are conversational features.
The ranges of the features are reported in Table 4.2, considering T=60 turns, and 77 different subjects.

We calculate statistical descriptors on them, that can be the mean values or the histograms for both gallery and probe set; in this last case, since the turns are usually short, we obtain histograms that are collapsed toward small numeric values.

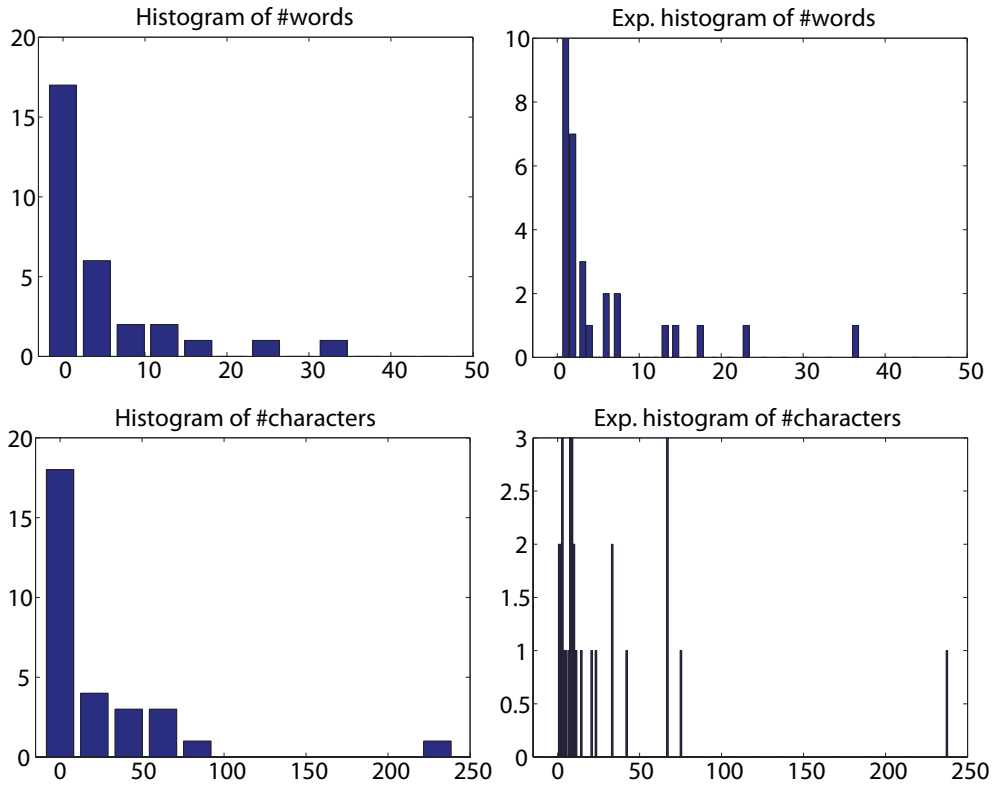


Figure 4.2: Distributions of some features: linear histogram (left), exponential histogram (right).

Modeling them as uniformly binned histograms over the whole range of the assumed values will produce ineffective quantizations, so we opt for exponential histograms, where small-sized bin ranges are located toward zero, increasing their sizes while going to higher numbers as shown in Figure 4.2.

The exponential histogram data structure is a histogram in which buckets recording older data are exponentially wider than the buckets recording more recent data. This intuition has been validated experimentally, as discussed in the following.

The introduction of turns as a basic analysis unit allows one to introduce features that explicitly take into account the conversational nature of the data and mirror behavioral measurements typically applied in automatic understanding of social interactions (see [41] for an extensive survey):

• **Turn duration**: the time spent to complete a turn (in hundredth of seconds); this feature accounts for the rhythm of the conversation with faster exchanges typically corresponding to higher engagement.

• **Writing speed** (two features): number of typed characters - or words - per second (typing rate); these two features indicate whether the duration of a turn is simply due to the amount of information typed (higher typing rates) or to cognitive load (low typing rate), i.e. to the need of thinking about what to write.

• **Number of "return" characters**: since these latter tend to provide interlocutors with an opportunity to start a new turn, high values of this feature are likely to measure the tendency to hold the floor and prevent others from "speaking" (an indirect measure of dominance).

• **Mimicry**: ratio between number of words in current turn and number of words in previous turn; this feature models the tendency of a subject to follow the conversation style of the interlocutor (at least for what concerns the length of the turns). The mimicry accounts for the social attitude of the subjects.

We call these features *conversational* features.

Table 4.2 provides basic facts about the features used in the experiments.

In the case of 1-13 and 16 the features correspond to the exponential histograms (32 bins) collected from the T turns. In the case of 14 and 15, the features correspond to the average estimated over the T turns. As we see, one feature has been

No.	Feature	Range
1	# words	[0,260]
2	# emoticons	[0,40]
3	# emoticons per word	[0,1]
4	# emoticons per characters	[0,0.5]
5	# exclamation marks	[0,12]
6	# question marks	[0,406]
7	# characters	[0,1318]
8	average word length	[0,20]
9	# three points	[0,34]
10	# uppercase letters	[0,94]
11	# uppercase letters/#words	[0,290]
12	**turn duration**	[0,1800(sec.)]
13	**# return chars**	[1,20]
14	**# chars per second**	[0,20(ch./sec.)]
15	**# words per second**	[0,260]
16	**mimicry degree**	[0,1115]

Table 4.2: Stylometric features used in the experiments. The symbol "#" stands for "number of". In bold, the conversational features.

removed because not much informative. This architectural choice maximized the AA accuracy.

For the remaining features, we estimate the mean, as we experimentally checked this was the best codification. One can note as some of the features are highly correlated, for example n.2,3,4, n.10,11 and n.14,15. Our aim is to instantiate a feature selection mechanism to select from these features which ones are the most informative, highlighting the underlying complementary.

Chapter 5

Experiments

Contents

The experiments have been performed over a corpus of dyadic chat conversations collected with Skype.
The conversations are spontaneous, i.e. they have been held by the subjects in their real life and not for the purpose of data collection.
This ensures that the behavior of the subjects is natural and no attempt has been made to modify the style in any sense.

The number of turns per subject ranges between 60 and 100. Hence, the experiments are performed over 60 turns of each person. In this way, any bias due to differences in the amount of available material should be avoided. When possible, we pick different turns selections (maintaining their chronological order) in order to generate different AA trials. The 60 turns of each subject are split into *probe* and *gallery* set, each including 30 samples. The average number of words per subject is 615.

5.1 Single feature CMC performance

The first part of the experiments aims at assessing each feature independently, as a simple ID signature.

A particular feature of a single subject is extracted from the probe set, and matched against the corresponding gallery features of all subjects, employing a given metrics (Bhattacharya distance for the histograms, Euclidean distance for the mean values). This happens for all the probe subjects, resulting in a $N \times N$ distance matrix. Ranking in ascending order the N distances for each probe element allows one to compute the *Cumulative Match Characteristic* (CMC) curve, i.e., the expectation of finding the correct match in the top n positions of the ranking.

In statistics, the Bhattacharyya distance measures the similarity of two discrete or continuous probability distributions. It is closely related to the Bhattacharyya coefficient which is a measure of the amount of overlap between two statistical samples or populations. The coefficient can be used to determine the relative closeness of the two samples being considered. It is used to measure the separability of classes in classification.

If we let R_i be the frequency coded quantity in bin i (normalised such that $\sum_i R_i = 1$) for the first histogram and S_i a similar quantity for the second histogram. Then we can assume R_i to be a Poisson distributed random variable and similarly for S_i.
The Bhattacharyya statistic $\sum_i \sqrt{R_i}\sqrt{S_i}$ can be a measure of similarity between the two histograms. For the case of two identical histograms we obtain $\sum_i R_i = 1$ indicating a perfect match.
The Bhattacharyya measure is: self consistent, unbiased and applicable to any distribution. The measure can be applied to the field of system identification [5].

CMC is a recognition rate that measure the probability that a given user appears in different sized candidate lists. The faster the CMC curve approaches 1, indicating that the user always appears in the candidate list of specified size, the better

the matching algorithm. CMC curve plots the probability of identification against the returned 1:N candidate list size.

It's an effective method of showing measured accuracy performance of AA approaches operating in the closed-set identification task [9].
Templates are compared and ranked based on their similarity. The CMC shows how often the individual's template appears in the ranks, based on the match rate. A CMC compares the rank versus identification rate.
In particular, the value of the CMC curve at position 1 is the probability that the probe ID signature of a subject is closer to the gallery ID signature of the same subject than to any other gallery ID signature; the value of the CMC curve at position n is the probability of finding the correct match in the first n ranked positions.

Given the CMC curve for each feature (obtained by averaging on all the available trials), the normalized Area Under Curve (nAUC) is calculated as a measure of accuracy.

Figure 5.1 shows that the individual performance of each feature is low (less than 10% at rank 1 of the CMC curve). In addition, the first dynamic feature (**Turn duration**) has the seventh higher nAUC, while the other ones are in position 10, 14, 15 and 16, respectively.

5.2 Our feature selection approach: forward feature selection

The experiments above serve as basis to apply the Forward Feature Selection (FFS) strategy.

Sequential forward selection (SFS) starts with an empty set and adds one feature at each step. The estimate of the quality of the candidate subsets usually depends on the training/testing split of the data.
The quality of a feature subset is measured by an estimate of the accuracy of the

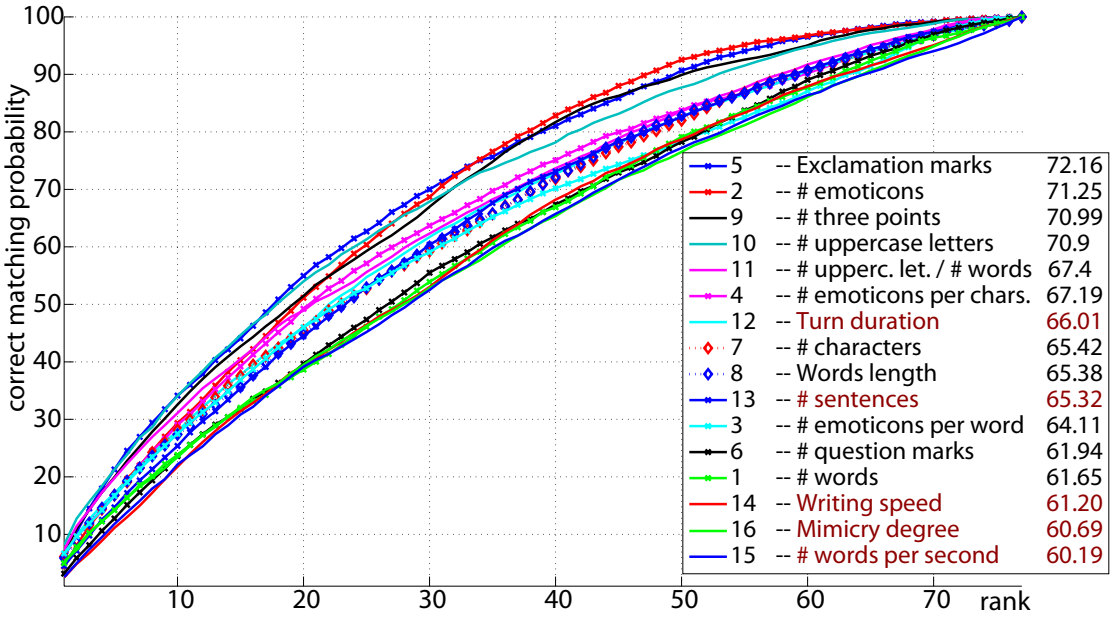


Figure 5.1: CMCs of the proposed features. The numbers on the right indicate the nAUC. Conversational features are in bold (best viewed in colors).

nAUC computed on the candidate subset.
Considering the 50 sequences obtained from the SFS runs, the stability index gives a measure of similarity between features.
Stability index keeps the most informative features (with high ranking in the SFS) that occurred most times. As minimum error of our SFS process we consider the maximum nAUC. For example, if a feature is at first position in the major of the 50 sequences, it will have a high similarity index, conversely if the same feature is presented in various position, it will have a low value of index. Thus aims to keep this feature as the one which has high performance and keep that sequences that satisfy this score, leaving out the others. In fact, indexes are based on cardinality of the intersection between the various sequences.

Forward Feature Selection (FFS) strategy serves to select the best pool of features that can compose an ID signature.
As stated Langley, four basic issues, that determine the nature of the heuristic search process, are reported.

• **Starting point:** The classic FFS is applied, so it starts with no features an successively add attributes. Each S_i is empty set.

• **Construction of candidate subsets:** A greedy approach is used. At each point in the search, one considers local changes to the current set of attributes, selects one, and then iterates, never reconsidering the choice. So, the FFS retains the feature with the highest nAUC, at the second one it selects the feature that, in combination with the previous one, gives the highest nAUC, and so on until all features have been processed. Combining features means to average their related distance matrices, forming a composite one. The pool of selected features is the one which gives the highest nAUC. Since FFS is a greedy strategy, different runs (50) of the feature selection are used, selecting a partially different pool of 30 turns each time for building the probe set. In this way, 50 different ranked subsets of features are obtained.

• **Criterion for evaluating the candidate subsets:** For distilling and evaluate a single subset, the Kuncheva stability index [24] is adopted. It gives us the best subset.

• **Stopping criterion:** We continue generating candidate sets until reaching the end of the search space (50 runs) and then select the best with stability index. The peak effect which normally occurs through the selection process suggests that there is an optimal number of features. The stability index is found with value 0.55 and the corresponding feature subset is returned. The optimal number of features is decided afterwards.

To find the best candidate feature sequence, we considered all features sequences from FFS process and compute the stability index.
Comparing with Kalousis Index [24] in Figure 5.2 we can notice that Kuncheva stability index is better.

In order to choose a final sequence of features, we find the sequence with the minimum local error, i.e the sequence with the highest nAUC. Having a set of sequences instead of a single one opens up a multitude of choices with respect to

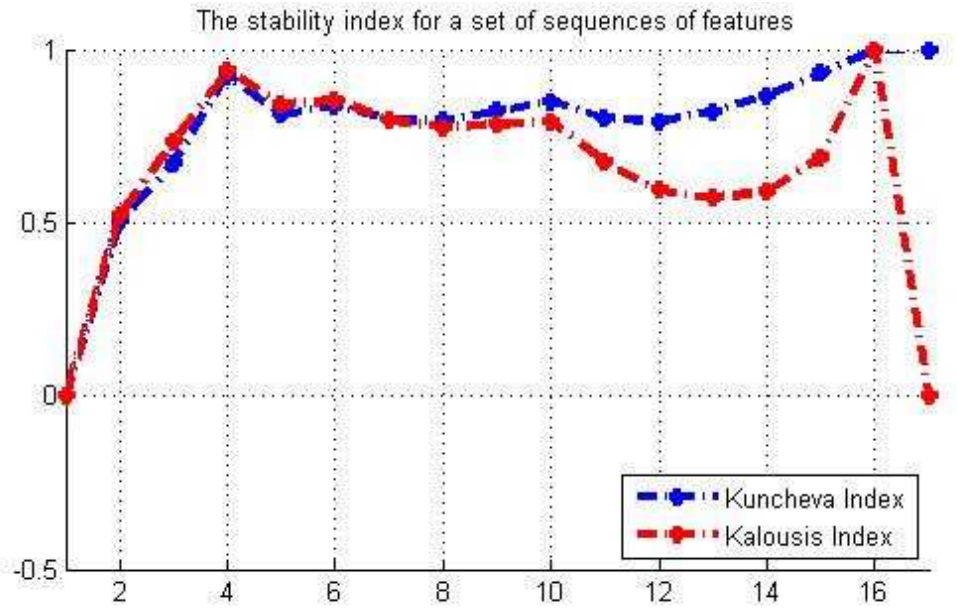


Figure 5.2: The stability index for a set of sequences of features

selecting the number of features d.
We find a set of "consistent values" d, for which the sequences agree i.e,

$$D_{cons} = d|1 \leq d \leq n, \mathcal{I}_S(\mathcal{A}(d)) > \theta,$$

where θ is a predefined threshold on $\mathcal{I}_S$. Choose d to be the one with the highest nAUC within D_{cons}. For all features we check if the AUC of d_i is better than the AUC considered with less features at previous step, and then we hold only the features that satisfy this option. If this sequence is also better than θ we found our best feature sequence.

According to $\mathcal{I}_S$ we opted to retain 12 features, since after this number, the stability index decreased.
The FFS process results into 12 features, ranked according to their contribution to the overall CMC curve. We obtain in this case a nAUC=90,53.
The set includes features in Table 5.1

No.	Feature
5	# exclamation marks
2	# emoticons
9	# three points
10	# uppercase letters
12	**turn duration**
13	**# return chars**
8	average word length
14	**# chars per second**
6	# question marks
7	# characters
15	**# words per second**
16	**mimicry degree**

Table 5.1: Final sequence of features.

In bold, we report the conversational features that appear to rank higher than when used individually. This suggests that, even if their individual nAUC was relatively low, they encode information complementary with respect to the traditional AA features.

The final CMC curve, obtained using the pool of selected features, is reported in Figure 5.3, curve (a). In this case, the rank 1 accuracy is 29.2%.
As comparison, other CMC curves are reported, considering (b) the whole pool of features (without feature selection); (c) the same as (b), but adopting linear histograms instead of exponential ones; (d) the selected features with exponential histograms, without the conversational ones; (e) the conversational features alone and (f) the selected features, calculating the mean statistics over the whole 30 turns, as done usually in the literature with the stylometric features.

Several facts can be inferred: our approach has the highest nAUC; feature selection improves the performance; exponential histograms work better than linear ones; conversational features increase the matching probability of around 10% in the first 10 ranks; conversational features alone give higher performance of standard stylometric features, calculated over the whole set of turns, and not over each one of them.

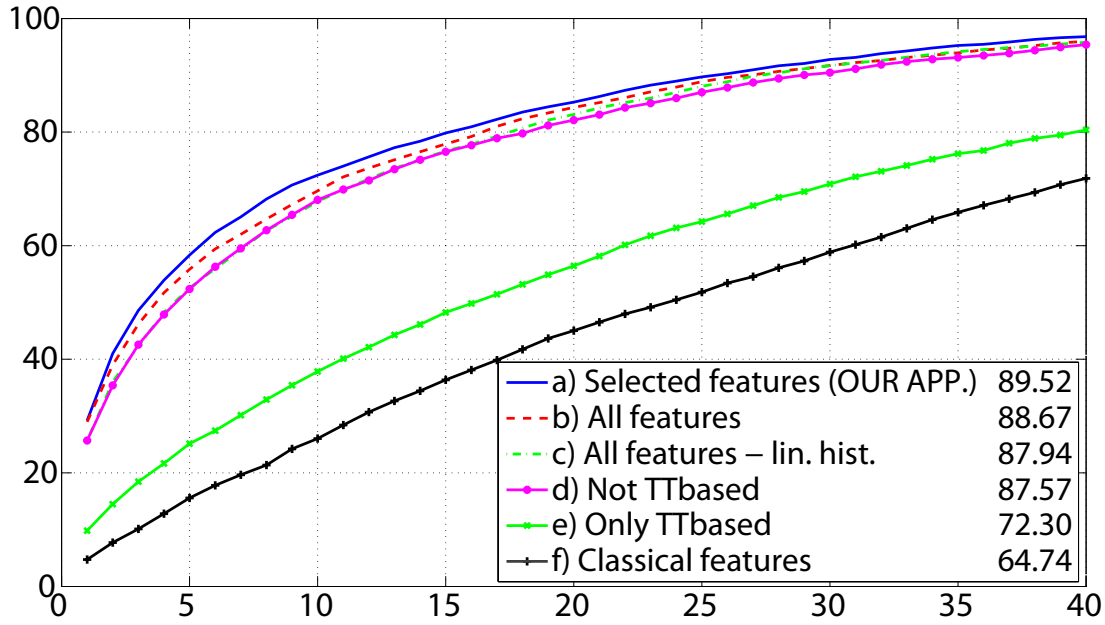


Figure 5.3: Comparison among different pool of features.

5.3 Relationship between performance and numbers of turns

The last experiment shows how the AA system behaves while diminishing the number of turns employed for creating the probe and gallery signatures. The results (mediated over 50 runs) are shown in Table 5.2.

# Turns	5	10	15	20	25	30
nAUC	68.6	76.6	80.6	85.0	88.4	89.5
rank1 acc.	7.1	14.0	15.1	21.9	30.6	29.2

Table 5.2: Relationship between performance and number of turns used to extract the ID signatures.

Increasing the number of turns increases the nAUC score, even if the the increase appears to be smaller around 30 turns.
In Figure 5.4 and Figure 5.5 shows how increase CMC and nAUC respectively to the growth of turns.

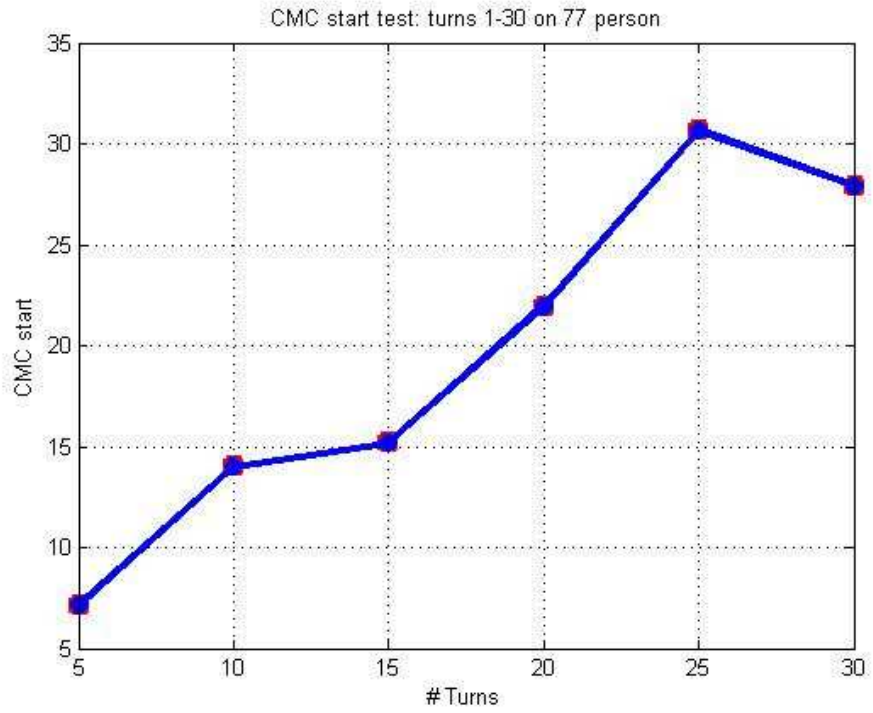


Figure 5.4: CMC performance as turns increase

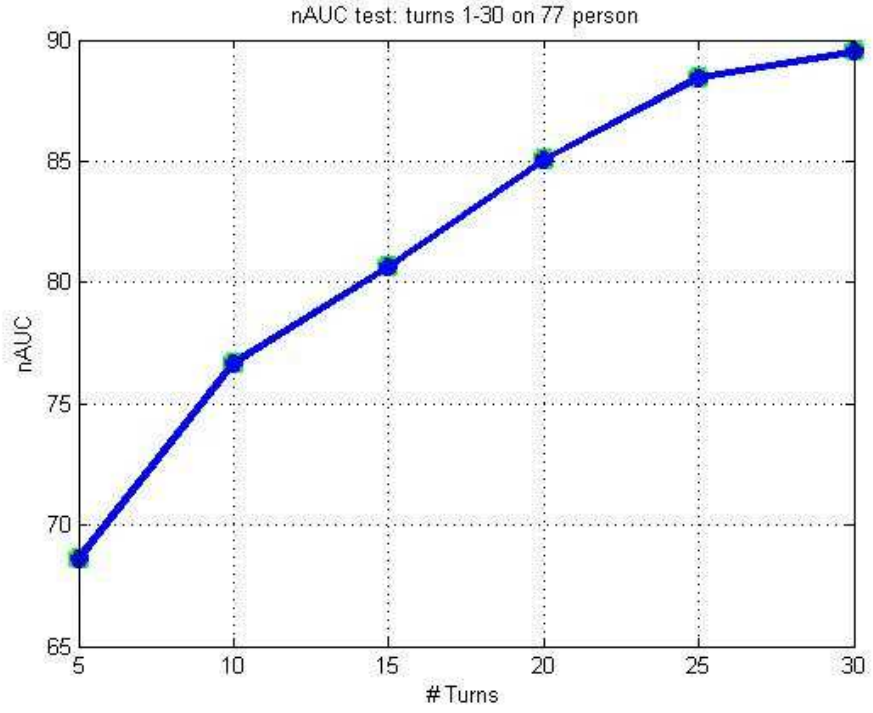


Figure 5.5: nAUC performance as turns increase

Conclusion

This study proposes two main contributions to the problem of recognizing automatically the identity of chat participants while respecting their privacy.
The first is the introduction of new features that account for turn-taking and mirror the features typically applied in automatic understanding of spoken conversations.
The second is the use of turns as a basic analysis unit for the analysis of chat data and identification of their participants.

The results are promising and show that taking into account the conversational nature of the texts typed during chat exchanges can improve the performance of AA approaches.

Many difficulties have been encountered. In the major part of SSP works, non-verbal cues are intended as non-verbal behavioral cues. We attempt to use those cues that are not associated to verbal conversation, but rather those which are related to chat conversation. Thus, we have introduced four novel feature: turn duration, writing speed, number of return characters and mimicry.

Another issue regards feature selection. We would like to seek a method that consider the information gain of each feature, to obtain the best feature set. It is addressed using the stability index, that allow to return the best features sequence and help us to find the best number of features.
To conduct a more accurate work, based on the value of the stability index, instead of consider the minimum local error, we had to take advantage of the rank of the features in each sequence.

A real novelty in this topic is the observation of the turns in a chat conversation. In all previous works no one have considered it, they have always treat entire chat conversation. The only studies that take into account turns don't regards AA but on seeking the behavior of people who deceives.

The strong hold lies in the CMC curve, that allows us to display clearly the results of the classification. Such curve shows that the selected features by the FFS process guarantees the best accuracy with AUC. We also have shown that the not conversational features, that we have introduced, provide an higher result than the only conversational ones.

Future Works

Future work will aim not only at continuing along such a direction, but also at adopting classifiers to improve the results; actually, in this approach no learning has been applied.

We can also consider the possibility of exploiting some works in speech analysis. The idea is to compare chat and speech conversation. This may lead some results on the similarity of people behavior between two different modes of conversation, always taking into account turn-taking.

'You must remember, family is often born of blood, but it doesn't depend on blood. Nor is it exclusive of friendship. Family members can be your best friends, you know. And best friends, whether or not they are related to you, can be your family.'

— Trenton Lee Stewart

Acknowledgments

In a few lines is hard to remember all those who in these years I have been close. Completing this master degree work has been a wonderful and often overwhelming experience.
It is hard to know whether it has been dealing with the social signalling itself which has been the real learning experience, or grappling with how to write a paper, work in a group, code intelligibly, stay up until the birds start singing, and . . . stay, um. . . focused.

I have been very privileged to have undoubtedly the most intuitive, smart and supportive advisor anyone could ask for, namely Marco Cristani.
I wish to thanks him for helping me in dealing with any critical situation of this project with confidence. He demonstrated the greater availability and he has been the cultural reference that every student would have. I'm also grateful for all the valuable advices and for most ideas generated during this theses and the possibility of being able to confront a person with his admirable ability.

I am also grateful to teachers and researchers who have supported me over the years of academic studies and who helped me during the thesis work, in particular dott.Marco Cristani, dott.Alessandro Vinciarelli, dott.Umberto Castellani , dott. Paolo Fiorini, dott. Federico Fontana and dott.Giorgio Roffo, who helped me during in nights spent fixing papers.

I would like to thank my parents for their love and support in all these years. Their help has been essential in order to reach this goal.

I thank and hug Alice for the demonstration of friendship and participation at any time, a true friend, and amazing colleague on work, always ready to watch movies and eat junk food all day with you!

A special hug goes my dearest friends who have made these five years memorable through theri comments and notes on my books. Ena, Giulia, Pic, Lora, Tome, Zeno, Saba, Mantovani, Soncini, Jasmine, Moz, Bru, Reekoz, Max, Omar, Monta, Samuel, Franci! Thanks for sharing with me the most challenges through and for allowing me to grow up beside you and always showing me great affection. A hug is just as special to Stefania for being at my side all these years, the escape valve of my Dionysian fury and despite everything, even the benchmark on which to rely. This achievement is yours. Thanks also for the beautiful summers spent together, remember Graziella, Gabriella and 'ntani'.

Thanks to colleagues Alice and Michy, working at Japan this past year.

Thanks Silvia, Ale, Dalco, Vale, Micky, Maik, Marco, Tome, Elena, Benny, Fede, Anto, Gess for the nights spent together and all the laughs of this year. Thanks Mattia, Renè, Dema, Pippo, Fabio, Cina, Ivan, Benati, Fava, Simon, Merz, Tome, Origin, Trivium, Dying Fetus, Psycroptic for good music, live concerts and funny time spent together.
Thanks to the intense study that in many days helped me to overcome bad times.
Thanks to the fake people in this world that allow you to better appreciate the people who love me and truly value the little things.
Thanks to Japan that has fed me for several months.

I was told to use a single page thanks, but I failed. A final special mention goes to all those who are not offended for not have been mentioned for first, second and third... but my affection in these years is grown so much that I have not forgotten you all. Remind you tonight as I was writing these pages that close another chapter of my life has been an immense joy for me!
For all of you a special " thank you thank you thank you!".

I would like to leave some conclusions on the experience at the University.

I had the privilege of dealing with passion and a hint of irony these challenges, which I shared with good friends and colleagues who have taught me how to improve myself.

I appreciate the importance of communication with the outside and within the team. Learning to listen and to express the right things.
I have worked hard to pass on my implicit and explicit knowledge, in oral and written, formal and informal with other project members.

In these years I had the opportunity to explore and learn technologies and tools that were not part of my academic curriculum.
I learned the importance of being involved and commit, as a means to gain the mutual respect of your colleagues. I learned to share with them difficulties and moments of tranquillity.

Verona July 17, 2012 C. S.

References

[1] A. Abbasi and H. Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.

[2] A. Abbasi and H. Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.

[3] A. Abbasi and H. Chen. Voiceprints: A stereometric approach to identity-level identification and similarity detection in cyberspace. *CAM Trans. Info. Sys.*, 26(2):1–29, 2008.

[4] A. Abbasi, H. Chen, and J. F. Nunamaker. Stereometric identification in electronic markets: Scalability and robustness. *Journal of Management Information Systems*, 25(1):49–78, 2008.

[5] F. J. Aherne, N. A. Thacker, and P. Rockett. The Bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 34(4):363–368, 1998.

[6] K. Albrecht. *Social Intelligence: The New Science of Success*. Jossey-Bass, A Wiley Imprint, 2006.

[7] N. Ali, M. Hindi, and R. V. Yampolskiy. Valuation of authorship attribution software on a chat bot corpus. In *Information, Communication and Automation Technologies*, pages 27–29. XXIII International Symposium on, 2011.

[8] S. Argamon, M. Koppel, J. Pennebaker, and J. Schler. Automatically profiling the author of an anonymous text. *Communications of CAM*, 52(2):119–123, 2009.

[9] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. *Guide to Biometrics*. Springer Verlag, 2003.

[10] K. Calix, M. Connors, D. Levy, H. Manzar, G. McCabe, and S. Westcott. Stylometry for e-mail author identification and authentication. *Interface*, 2008.

[11] A. Castro, O. Sotoye, L. T. Torres, V. Monaco, and J. Stewart. A stylometry system for authenticating students taking online tests. In *Proceedings of The Michael L. Gargano 9th Annual Student/Faculty Research Day*. NY, USA, 2011.

[12] M. Corney. Analysing e-mail text authorship for forensic purposes. Master's thesis, Queensland University of Technology, 2003.

[13] O. De Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *CAM SIGMOD Record*, 30(4), 2001.

[14] M. Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the International Conference on Computational Linguistics*, page 611, 2004.

[15] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting personality from twitter. In *SocialCom/PASSAT*, pages 149–156. IEEE, 2011.

[16] J. Goldstein-Stewart, R. Winder, and R. E. Sabin. Person identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 336–344. Association for Computational Linguistics, 2009.

[17] R. Goodman, M. Hahn, M. Marella, C. Ojar, and S. Westcott. The use of stylometry for email author identification : A feasibility study. *Proceedings of Student Faculty Research Day Pace University*, pages 1–7, 2007.

[18] D. I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998.

[19] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 2011.

[20] M. Koppel, S. Argamon, and A. R. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[21] P. V. Kranenburg and E. Backer. Musical style recognition - a quantitative approach. In *Proceedings of the Conference on Interdisciplinary Musicology*, 2004.

[22] I. Krsul and E. H. Spafford. Authorship analysis: Identifying the author of a program. In *Proc. 18th NIST-NCSC National Information Systems Security Conference*, pages 514–524, 1996.

[23] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can. Chat mining: Predicting user and message attributes in computer-mediated communication. *Info. Process. Manage.*, 44(4):1448–1466, 2008.

[24] L. Kuncheva. A stability index for feature selection. In *IASTED International Multi-Conference Artificial Intelligence and Applications*, pages 390–395, 2007.

[25] M. D. Levine. Feature extraction: A survey. *Proceedings of the IEEE*, 57(8):1391–1407, 1969.

[26] P. M. Mccarthy, G. A. Lewis, D. F. Dufty, and D. S. Mcnamara. Analyzing writing styles with coh-metrix. *Methods*, (1995):764–769, 2004.

[27] H. Mohtasseb and A. Ahmed. Mining online diaries for blogger identification. *Learning*, 2009.

[28] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, R. Shin, and D. Song. On the feasibility of internet-scale author identification. In *Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy*, pages 22–28. Westin St.Francis, San Francisco, CA, IEEE, 2012.

[29] A. Orebaugh. *An Instant Messaging Intrusion Detection System Framework: Using character frequency analysis for authorship identification and validation*, pages 160–172. 2006.

[30] A. Orebaugh and J. Allnutt. Classification of instant messaging communications for forensics analysis. *Social Networks*, pages 22–28, 2009.

[31] D. Pavelec, E. J. R. Justino, and L. S. Oliveira. Author identification using stereometric features. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, 11(36):59–66, 2007.

[32] J. Resig, S. Dawara, C. M. Homan, and A. Teredesai. Extracting social networks from instant messaging populations. *Social Networks*, (Im), 2004.

[33] J. Resig and A. Teredesai. A framework for mining instant messaging services, 2004.

[34] G. Shalhoub, R. Simon, R. Iyer, J. Tailor, and S. Westcott. Stylometry system - use cases and feasibility study. In *Proceedings of Student-Faculty Research Day*, pages A3.1–A3.8, 2010.

[35] E. H. Spafford and S. A. Weeber. Software forensics: Tracking code to its authors. *Computers and Society*, 1993.

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

[37] E. Svoboda. Is that painting real? ask a mathematician. USA Today, May 2007.

[38] E. Tupes and R. E. Christal. Recurrent personality factors based on trait ratings. Technical report, Lackland Air Force Base, TX: US Air Force, 1961.

[39] F. J. Tweedie, S. Singh, and D. I. Holmes. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, 30(1):1–10, 1996.

[40] B. Vickers. Shakespeare and authorship studies in the twenty-first century. *Shakespeare Quarterly*, 62(1):106–142, 2011.

[41] A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.

[42] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signal processing: state-of-the-art and future perspectives of an emerging domain. In *Proceedings of the 16th CAM international conference on Multimedia*, MM '08, pages 1061–1070. CAM, 2008.

[43] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, 2006.

[44] D. Zhou, L. Zhang. Can online behavior unveil deceivers? an exploratory investigation of deception in instant messaging. In *Proceedings of the Hawaii International Conference on System Sciences*, 2004.

Colophon

Note: This thesis was typeset using LaTeX.
The body text is set in 12 point Computer Modern Roman and issued by Foxit Reader in 2012.

Final Version: *July 17, 2012.*

Statement

I declare that this research is my own work, except for those parts explicitly mentioned in the text, and that this work was not proposed for any academic or professional qualification, except as explicitly stated.

I also certify that the material contained in this work can be published, whole or in part, from my Master Degree thesis advisor in Engineering and Computer Science, Prof. Marco Cristani, of the University of Verona.

Verona July 17, 2012

Cristina Segalin