RedFame

Studies in Media and Communication
Vol. 6, No. 2; December 2018
ISSN: 2325-8071 E-ISSN: 2325-808X
Published by Redfame Publishing
URL: http://smc.redfame.com

Evaluation of the Performance and Efficiency of the Automated Linguistic Features for Author Identification in Short Text Messages Using Different Variable Selection Techniques

Refat Aljumily

Correspondence: Refat Aljumily, Freelance Forensic Linguist and Data analyst, United Kingdom.

Received: November 25, 2018 Accepted: December 18, 2018 Online Published: December 20, 2018

doi:10.11114/smc.v6i2.3892 URL: https://doi.org/10.11114/smc.v6i2.3892

Abstract

The aim of this paper was to evaluate the efficiency of automated linguistic features to test its capacity or discriminating power as style markers for author identification in short text messages of the Facebook genre. The corpus used to evaluate the automated linguistics features was compiled from 221 Facebook texts (each text is about 2 to 3 lines/35-40 words) written in English, which were written in the same genre and topic and posted in the same year group, totaling 7530 words. To compose the dataset for linguistic features performance or evaluation, frequency values were collected from 16 linguistic feature types involving parts of speech, function words, word bigrams, character tri grams, average sentence length in terms of words, average sentence length in terms of characters, Yule's K measure, Simpson's D measure, average words length, FW/CW ratio, average characters, content specific key words, type/token ratio, total number of short words less than four characters, contractions, and total number of characters in words which were selected from five corpora, totalling 328 test features. The evaluation of the 16 linguistic feature types differ from those of other analyses because the study used different variable selection methods including feature type frequency, variance, term frequency/ inverse document frequency (TF.IDF), signal-noise ratio, and Poisson term distribution. The relationships between known and anonymous text messages were examined using hierarchical linear and non-hierarchical nonlinear clustering methods, taking into accounts the nonlinear patterns among the data. There were similarities between the anonymous text messages and the authors of the non-anonymous text messages in terms function word and parts of speech usages based on TF.IDF technique and the efficiency of function word usages (=60%) and the efficiency of parts of speech frequencies (=50%). There were no similarities between the anonymous text messages and the authors of the non-anonymous text messages in terms of the other features using feature type frequency and variance techniques in this test and the efficiency of these features in the corpus (< 40%). There was a positive effect on identification performance using parts of speech and function word frequency usages and applying TF.IDF technique as the length of text messages increased (N≥ 100). Through this way, the performance and efficiency of syntactic features and function word usages to identify anonymous authors or text messages is improved by increasing the length of the text messages using TF.IDF variable selection technique, but decreased as feature type frequency and variance techniques in the selection process apply.

Keywords: stylometry, linguistic features, hierarchical linear clustering, non-hierarchical non-linear clustering, distance metrics, variance, signal-noise ratio, poisson frequency distribution,TF.IDF term-frequency, SOM

1. Research Problem

What linguistic feature type(s) can successfully be used to determine if two anonymous short text messages are written by the same author?

2. Introduction

What linguistic feature type(s) should we use to create a profile that characterizes the writing style of an author in a short text message? Well in real-world applications it is not always clear what linguistic features are appropriate in a given case or what linguistic features will help us model the problem we are trying to solve. Along with this problem, there is also another problem with linguistic features being unimportant or not being very useful. Evaluating the performance and efficiency of linguistic features is a systematic process that- if done correctly-can lead to obtain valid information for a research problem on what linguistic features to observe, what is important, and what is not. Feature selection methods are techniques used to help automatically pick out important linguistic features. There are many articles written on linguistic features that entirely ignore this part of the process. The aim is to evaluate the efficiency and performance of linguistic features with respect to stylistic variation (i.e. author stylistic preferences and linguistic choices), genre variation, topic variation, and various length portions of texts. The intention of this study is to evaluate the performance of such features. Specifically, the goals are:

1. To test its capacity or discriminating power as style markers for author identification in short text messages of specific text genres (i.e. Facebook genre), where a short text message is defined here as any short communication unit about 2 to 3 lines long that falls somewhere between talking and writing or writing more freely with less anxiety about audience including email, Facebook comment, Tweet feed, short online messages, writing in chat rooms that are used in any social networking sites or websites.
2. To use different variable selection techniques and see which feature selection technique finds the best possible linguistic features for the desired pattern or an optimally relevant set of linguistic features for a research question.
3. To show how the short texts impact on the identification performance and feature selection techniques.

3. Linguistic Features and Feature Selection Techniques in Stylometry

There are many linguistic features that can be used in stylometric authorship attribution and can be classified into character based, word-based, sentence-based, structural or syntactic features. Each of these features can be sub-classified into so many different types. For example:

1. Character-based features: This feature includes letters, n-character sequences, n-grams, etc. A character n-gram is defined as a string of contiguous alphanumeric symbols, perhaps including also punctuation symbols and contractions.
2. Word-based: This feature is commonly applied in different ways:
   - Average word length, word-length frequency distributions, the number of syllables per word, and the number of phonemes per word. The length of a word is defined as the number of letters which constitute it.
   - Vocabulary richness: the degree of diversity of vocabulary in a text.
   - Function words: such as pronouns, auxiliary verbs, prepositions, conjunctions, determiners, degree adverbs, negations, quantifiers, and relativizers.
   - Content words: such as nouns, verbs, adjectives, and adverbs
   - Word n-grams: A word *n*-gram is defined as a string of words, where each *n*-gram is composed of *n* words.
   - Syntactic or structural features: such as phrasal composition grammar, distribution of parts of speech and re-write rules; (preferred) word position, etc.

These and many other linguistic features have been measured and analyzed by Stylometry: the science of measuring linguistic style which can be used to determine authorship from various linguistic features contained within the texts themselves. Stylometry in general assumes that one part of an author’s choice of linguistic features is conscious, deliberate, and open to imitation or borrowing by others. The other is sub-conscious, that is, independent of an author’s direct control, and far less open to imitation or borrowing. Stylometry focuses on the unconscious part of an author’s choice of linguistic features and assumes that at least some aspects of it are constant across his or her literary output, and suggests that these constants can be identified and applied to areas like authorship attribution on the basis of quantitative criteria using computational methods. This assumption is very well-known and well-understood and explained, and there is an extremely long line of research on this subject (e.g. Aljumily, 2015; 2017; Holmes, 1994, 1998; Holmes and Kardos, 2003). More details about the linguistic features for authorship attribution that have been proposed in the literature is discussed in, for example, Stamatatos (2009).

Similarly, just as there are many linguistic features that the authors of disputed or anonymous texts can be identified, there are many techniques, or methods, by which the best possible linguistic feature or set of linguistic features (i.e. features with high or predictive clues) can be selected using their frequency to build a model. Feature selection is a crucial issue in many classification/clustering, and characterization problems especially when the learning task involves high-dimensional dataset, and the remainder of this section provides a bit of information about the ones used most often used in authorship tests.

1. Feature selection based on lexical type frequency

Lexical type frequency is a simple technique to feature selection. It records only the presence or absence of the value of any lexical type in a text which can be weighted by multiplying the value by the frequency of the type in text. It selects the most useful features in a corpus as key markers, and discarding the less useful ones, as expressed by the equation:

$$Q_{ij} = Q_{ij} freq_{ij} \tag{1}$$

By default, it selects a maximum frequency threshold, removes all features with frequencies greater than the threshold, and retains the medium-frequency features between the maximum and minimum thresholds as the key markers for distinguishing between and among texts (van Rijsbergen 1979, Belew, 2000).

2. Feature selection based on variance

Variance threshold is another simple frequency technique to feature selection. The variance of a set of feature values is the average deviation of those values from their mean, as expressed by the equation:

$$v = (\sum_{i=1..n} (x_i - \mu)^2) / n \tag{2}$$

By default, it removes all features with low variance, i.e. features whose values do not vary enough for them to be useful in text analysis (Pyle 1999). Variance is a fundamental concept in probability and statistics, and any in a wide range of textbooks can provide additional information --see, for example, Milton & Arnold, (2003).

3. Feature selection based on Lexical frequency distribution

This technique suggests that a feature type's value is obtained not by its absolute frequency across a text, but by the pattern of variation in its frequency of occurrence across the texts. The three most often used techniques for obtaining such patterns of occurrence are:

- TF-IDF (Term Frequency-Inverse document frequency)

This technique is the product of term frequency (TF) and inverse document frequency (IDF), which is a weight often used to calculate pattern of occurrence of feature types based on their importance:

$$TF.IDF_{mT} = f_m \, log2 \, (n_T / tf_m) \tag{3}$$

where

- T is a text corpus
- f_m is the frequency of m across all texts in T
- IDF$_{mT}$ is the inverse text frequency of feature type m across the total number of texts in T
- n_T is the total number of texts in T
- tf_m is the number of texts in T that contain m
- $\log_2$ is not conceptually part of IDF, but merely scales n_T / tf_m to a convenient numerical interval.

This technique evaluates how important a feature type is to a text in a corpus. The importance increases proportionally to the number of times feature occurs in the text but is offset by the occurrence frequency of the feature in the corpus. More specifically, this technique is consisted by two terms: the first calculates the Term Frequency (TF) (the number of times a feature type appears in text, divided by the total number of feature types in that text; the second term is the Inverse Document Frequency (IDF), calculated as the logarithm of the number of texts in the corpus divided by the number of texts where the particular feature type occurs. By default, this technique calculates the TF.IDF for each column of data matrix and retains only those columns with a TF.IDF above a specified threshold (Spärck-Jones, 2004; 1974, 1972).

- Signal-noise ratio

Signal-noise ratio is a fundamental part of the information theory that compares the level of a desired signal to the level of background noise. Signal-noise ratio refers to the ratio of signal power (signal sequences or symbols or information) to the level of noise (entropy). A ratio higher than 1: 1 indicates more signal than noise. This idea is often used to reduce dimensionality by referring to the ratio of useful information to redundant or irrelevant data in a conversation or exchange, and thus becomes closely related to identification of ‘clumpy lexical type’ occurrence across a text corpus. The higher the entropy of a column vector in the frequency matrix, the more equally distributed the frequencies are, and thus the less important the lexical type connected with that column vector is for discriminating the texts in the corpus; on the contrary, the lower the entropy, the more clumpy the distribution, and the more important the lexical type. This can be calculated from:

$$signal_m = log2(totalfrequency_m) - noise_m \tag{4}$$

where $totalfrequency_m$ is the sum of frequencies in column vector m. The signals for all the columns in the matrix are calculated and sorted into descending order of magnitude and all the column vectors whose signals don’t meet a specified threshold are removed. (Salton & McGill, 1983, Belew, 2000).

- Poisson term probability distribution

In probability theory and statistics, the Poisson distribution is basically used to express the probability of a given number of random and rare events occurring in a specified spatial or temporal period of time, and is expressed as:

$$P(X = r) = (e^{-\lambda}\lambda^{r}) / r! \tag{5}$$

where:

- The symbol P stands for 'probability'
- *X* is a random variable
- *r* is the number of events that occur over a period i
- e is the base of the natural logarithm 2.71828...
- λ is the average value of *X* over many periods i

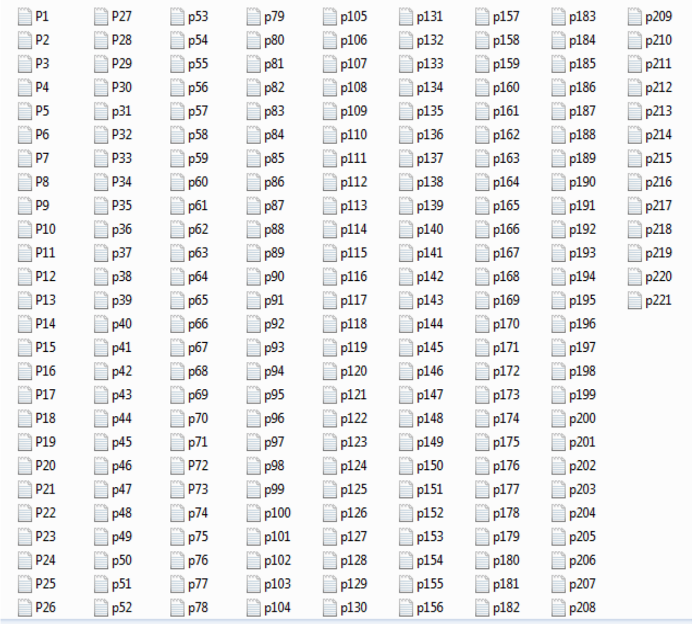
The Poisson distribution can be useful to select feature types by evaluating the patterns of clumpy of lexical type occurrence in a text corpus by the degree to which feature types are non-randomly distributed across the texts: on the one hand, a feature type that is randomly distributed across all the texts is taken to be unimportant for classification and can be removed, and on the other a feature type that has a non-random pattern of occurrence, that is, once that occurs frequently in a relatively small subset of texts and little or not at all in others, is taken to be an important classification criterion and is kept as a style marker. The application of Poisson term probability distribution to dimensionality reduction is straightforward: measure and calculate the differences between mean and variance for all features in a corpus and eliminate features whose frequency value doesn’t meet some threshold value. Church and Gale (1995a, 1995b), Belew (2000).

That concludes the background information for this section. It is important to be aware that there is much more to say about feature selection techniques in terms of their algorithms, advantages and disadvantages. Stay tuned for the next article on the subject of feature selection techniques, in which I will discuss them in more detail.

4. Methodology

4.1 Data

For the needs of the present research I developed a corpus of 221 Facebook texts in English and from the same year group, namely: 211 texts were status posts written by 211 users. These acted as the known texts. Another 10 texts were comments written by 10 of those 211 users who each commented on a post about women’s superiority over men. These texts acted as the unknown texts. The underlying rationale of using this genre is to investigate authorship attribution in the context of short online messages and to test the practical applicability of Stylometry in this form of E.writing, which most users see as a communication that falls somewhere between and writing or writing more freely with less anxiety about audience The criterion for selecting the users was the topic of the available Facebook text samples. This was to (1) minimize the topic factor in distinguishing among the text samples and to (2) reduce the impact of genre factor. All text samples were converted into a machine-readable format and coded/anonymised according to date and time. The text samples messages were distinguished by the letter P and a number following each code. For example, P (10) refers to person number 10 in the corpus**.** Altogether, therefore, there were 221 text samples, given as follows:



P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16
P17
P18
P19
P20
P21
P22
P23
P24
P25
P26
P27
P28
P29
P30
p31
P32
P33
P34
P35
p36
p37
p38
p39
p40
p41
p42
p43
p44
p45
p46
p47
p48
p49
p50
p51
p52
p53
p54
p55
p56
p57
p58
p59
p60
p61
p62
p63
p64
p65
p66
p67
p68
p69
p70
p71
P72
P73
p74
p75
p76
p77
p78
p79
p80
p81
p82
p83
p84
p85
p86
p87
p88
p89
p90
p91
p92
p93
p94
p95
p96
p97
p98
p99
p100
p101
p102
p103
p104
p105
p106
p107
p108
p109
p110
p111
p112
p113
p114
p115
p116
p117
p118
p119
p120
p121
p122
p123
p124
p125
p126
p127
p128
p129
p130
p131
p132
p133
p134
p135
p136
p137
p138
p139
p140
p141
p142
p143
p144
p145
p146
p147
p148
p149
p150
p151
p152
p153
p154
p155
p156
p157
p158
p159
p160
p161
p162
p163
p164
p165
p166
p167
p168
p169
p170
p171
p172
p173
p174
p175
p176
p177
p178
p179
p180
p181
p182
p183
p184
p185
p186
p187
p188
p189
p190
p191
p192
p193
p194
p195
p196
p197
p198
p199
p200
p201
p202
p203
p204
p205
p206
p207
p208
p209
p210
p211
p212
p213
p214
p215
p216
p217
p218
p219
p220
p221

Figure 1. The created corpus of 221 text messages

The 20 texts, 2 texts per Facebook user, for 10 Facebook users will be referred to as texts of interest and be classified into known and anonymous. These texts are presented in Table (1).

Persons	Known texts	Anonymous texts
1	P202	p203
2	P45	P216
3	P179	P217
4	P205	P211
5	P209	P214
6	P207	P212
7	P178	P219
8	P208	P213
9	P210	P204
10	P206	P220

The corpus consisted of totals around 7530 words. The number of words per text sample was between 30 and 35 words in length and the number of lines per email was between 2 and 3 in length. The corpus was preprocessed prior to analyzing it. All identifiers I was not interested in were deleted from the created corpus. This included punctuation marks, alphabet letters, special symbols, digits, and other unnecessary marks. All the comments and posts were filtered out as well keeping just the plain texts posted by the users thus reducing it from 7530 to 2792 words.

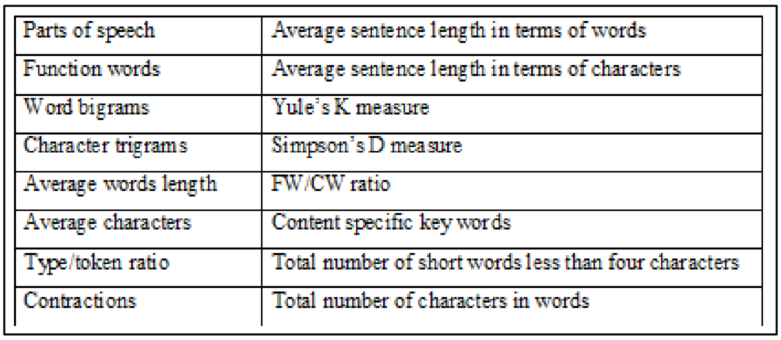
4.1.1 Rationale for the Very Short Text Messages

It is known that forensic stylometric methods can achieve high accuracy rates for long texts but they can achieve low accuracy rates for short texts, in particular when dealing with large number of authors. This is a very difficult task because short texts contain so little measurable style identifiers and/or so little repeated information. This behaviour motivates me to conduct and test the current methodology on short online text messages by a large number of authors. Given that text samples used are very short in length, the test results not only contribute to evaluate the usefulness of style identifiers used here for authorship attribution of online text messages, but also allow analysing their appropriateness to handle difficult attribution scenarios.

4.1.2 Stylometric Identifiers

For this test I began by thinking to measure almost anything in the created corpus- but I changed my mind that not only does any such measurement have to be relevant, it should have some basis in theory too. I wondered, therefore, what categories of word and what syntactic identifiers could be measured which would have some rationale in authorship attribution. Reverting to ideas formed in forensic styolmerty theory I considered the hypothesis that function words are stylo-linguistically neutral and (relatively) unconscious. (Grieve, 2007; Argamon and Levitan, 2005**)**. I also looked at the hypothesis that syntactic identifiers carry information which could be useful for stylometric analysis to capture the distinctive aspects of someone's writing style. I decided to count the number of different parts of speech such as nouns, verbs, adjectives, adverbs and other POS tags in a text which can also be useful stylistic characteristics in distinguishing between authors. Further, I considered word and character bigrams for being useful in attributing authorship. The reason why character n-grams provide good clues to authorship could be that they capture and combine information on different linguistic levels: lexical, syntactic, and structure (Houvardas and Stamatatos, 2006). I then thought about using the linguistic identifiers that appear inside words themselves such as average word length, contractions, vocabulary richness, and lexical density, the identifiers that appear within words themselves such as average sentence length. Most of these identifiers are considered unreliable when used by themselves, but I used them to examine their behavior in the short-length texts corpus. The aim was to determine which style identifiers make it possible to correctly identify the writer of the unknown text messages. Altogether, therefore, there were 16 style identifiers, given as follows:

Table 2. Style identifiers



Parts of speech	Average sentence length in terms of words
Function words	Average sentence length in terms of characters
Word bigrams	Yule's K measure
Character trigrams	Simpson's D measure
Average words length	FW/CW ratio
Average characters	Content specific key words
Type/token ratio	Total number of short words less than four characters
Contractions	Total number of characters in words

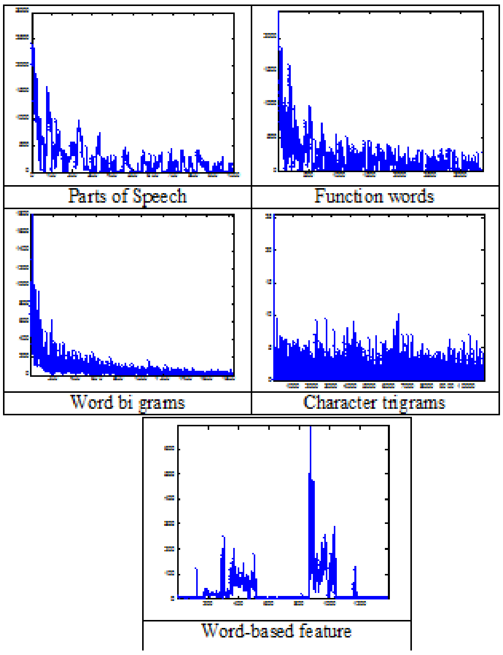
Each identifier was computed for its value in each of the 221 texts, giving 2279 function words, 1800 word bigrams, 5989 letter trigrams, and 11 word-based features. As for POS features, one kind of syntactic identifiers was defined: part-of speech tag frequencies. I used a Stanford POS tagger to parse the 221 texts and produce the parts-of speech for each text, giving 2788 parts of speech. These POS tags were used to match texts and identify similarities in structures.

4.1.3 Data Representation and Adjustment

The 221 text messages were converted into 221 vectors in a high dimensional space, and the 2279 function words, 1800 word bi-grams, 5989 letter tri-grams, and 11 word-based identifiers counted in the corpus were stored in four data matrices:, 221 × 2279 D_{FW}, 221 × 1800 $D_{bigram,}$ 221 × 8995 $D_{trigram}$, and 221 x 11 $_{\text{Dword-based}}$. The frequency values for D_{POS} were obtained by counting, for each text, the number of times each of the POS tag occurs, yielding a data matrix of 221 x 2788. The five data matrices were adjusted for (i) length, (ii) input dimensionality, (iii) variable scales. In spite of being balanced in terms of the size of words and the number of lines (as described above), four generated data matrices (Dpos, D_{FW}, $D_{wordbigrams}$, and $D_{chtrigrams}$) were adjusted for lengths. Texts from one section of the corpus may be much longer than those in another and may lead to conclude that a particular stylometric identifier is much more common in a particular author or text, where proportionally this is not actually the case. To avoid this possibility, in each row vector, the count for a given variable was multiplied by the mean document length, then divided by the total number of frequency counts occurring in that row vector. This normalization was relative to mean text length across a collection. The fifth data matrix (i.e. Dword-based) was column vector standardized to take the effect of variation in scaling among variables and to make each variable receives equal weight in the cluster analysis. To achieve this, in each column vector, the value of a given numerical column vector in the unstandardized matrix was divided the mean of column vectors. This standardization was relative to standardization based on variable mean.

The five data matrices (Dpos, D_{FW}, $D_{wordbigrams}$, $D_{chtrigrams}$, and $D_{word\text{-}based}$) were also adjusted for input dimensionality to select a suitable set of variables from the original identifiers in each data matrix. The POS, FW, W.bi-gram, and Char.bi-tri-gram data matrices had a large number of variables while word-based data matrix had relatively a few variables but there was scope for dimensionality reduction. Five different selection methods that are algorithmically different from each other were used to select the most discriminative identifiers for authorship identification test: lexical type frequency, variance, TF-IDF, signal-ratio noise, and Poisson term frequency distribution. TF-IDF was the best criterion for text messages identification, and therefore was selected in the current application. TF-IDF was chosen not by chance or because I preferred to use only that method, but was chosen because it maximized feature relevancy and reduced feature redundancy and therefore improved accuracy by giving higher weights to identifiers used by fewer authors. That is, an identifier that occurs, for example, 5 times in a single text message thereby has a TF.IDF five times as large as an identifier that occurs only once, which is both intuitively satisfying and also eliminating the effect of all very infrequent identifiers on text clustering.

Applying TF.IDF on the data matrices, a set of the highest POS, FWs, W.bi-grams, Char.bi-tri-grams, and words based identifiers score textfiles were selected as profile POS/FW/W.bi-gram/Char.bi-tri-gram summary. The TF.IDF for each column of data matrix was calculated and saved. The columns of data matrix with a TF.IDF above a specified threshold was only retained. And a suitable threshold for distinguishing between and among texts was selected, as shown in Figure (3):



Parts of Speech

Function words

Word bi grams

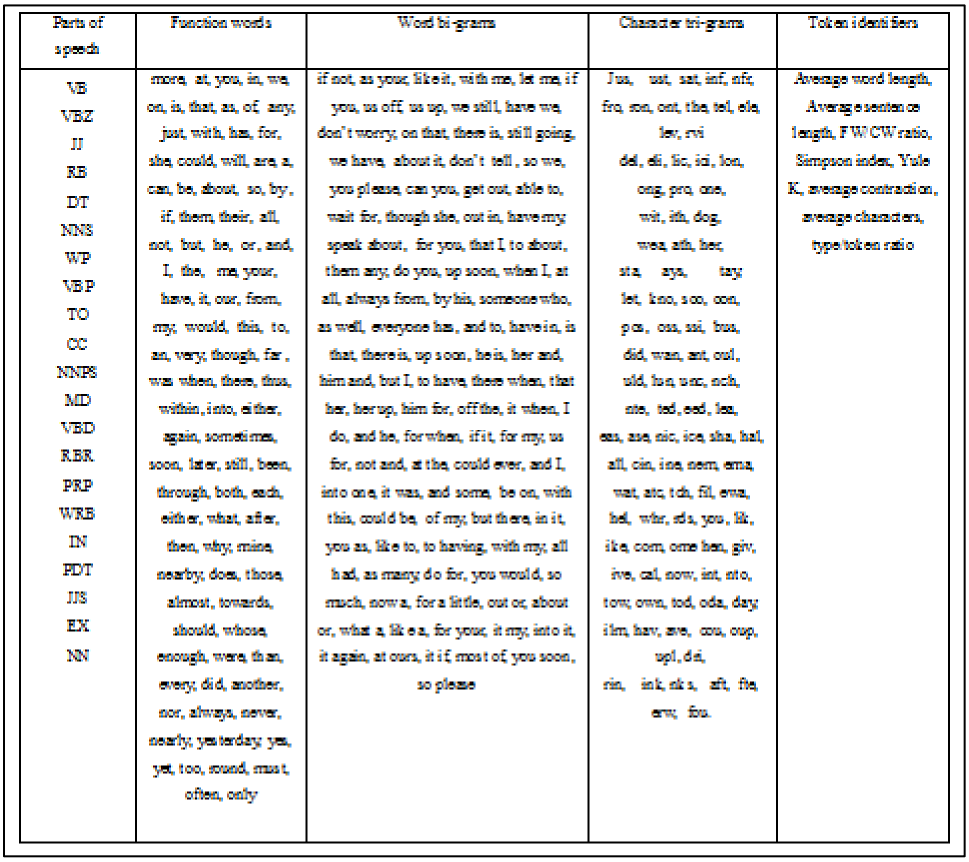
Character trigrams

Word-based feature

Figure 3. The selection of the highest TF.IDF identifiers

D_{POS} adjusted from a 221 x 2278 into a 221 x 21, D_{FW} transformed from a 221 x 2279 into a 221 x 100; D_{bigram} transformed from a 221 x 1800 into a 221 x 100; $D_{trigram}$ transformed from a 221 x 8995 into a 221 x 100; and finally D_{token} transformed from a 221 x 11 words into a 221 x 7. 100 function words, 1800 word bi-grams, 100 letter tri-grams, and 7 token counted in the corpus were stored in these vectors. A 221 x 21 D_{POS}, 221 x 2278, 221 × 100 D_{FW}, 221 × 100 D_{bigram}, 221 × 100 $D_{trigram}$, and 221 x 7 data matrices were computationally generated. The selected identifiers represented the linguistic profile of the participants, and these are presented in Table (3):

Table 3. The selected identifiers 21 parts of speech, 100 function words, 100 word bi-grams, 100 character tri grams, and token based identifiers with the highest TF.IDF



Parts of speech	Function words	Word bi-grams	Character tri-grams	Token identifiers
VB VBZ JJ RB DT NNS WP VBP TO CC NNPS MD VBD RBR PRP WRB IN PDT JJS EX NN	more, at, you, in, we, on, is, that, as, of, any, just, with, has, for, she, could, will, are, a, can, be, about, so, by, if, them, their, all, not, but, he, or, and, I, the, me, your, have, it, our, from, my, would, this, to, an, very, though, far, was when, there, thus, within, into, either, again, sometimes, soon, later, still, been, through, both, each, either, what, after, then, why, mine, nearby, does, those, almost, towards, should, whose, enough, were, than, every, did, another, nor, always, never, nearly, yesterday, yes, yet, too, round, must, often, only	if not, as your, like it, with me, let me, if you, us off, us up, we still, have we, don't worry, on that, there is, still going, we have, about it, don't tell, so we, you please, can you, get out, able to, wait for, though she, out in, have my speak about, for you, that I, to about, them any, do you, up soon, when I, at all, always from, by his, someone who, as well, everyone has, and to, have in, is that, there is, up soon, he is, her and, him and, but I, to have, there when, that her, her up, him for, off the, it when, I do, and he, for when, if it, for my, us for, not and, at the, could ever, and I, into one, it was, and some, be on, with this, could be, of my, but there, in it, you as, like to, to having, with my, all had, as many do for, you would, so much, now a, for a little, out or, about or, what a, like a, for your, it my, into it, it again, at ours, it if, most of, you soon, so please	Jus, ust, sat, inf, nfr, fro, son, ont, the, tel, ele, lev, rvi del, eli, lic, ici, lon, ong, pro, one, wit, ith, dog, wea, ath, her, sta, ays, tay, let, kno, soo, oon, pos, oss, ssi, bus, did, wan, ant, oul, uld, lun, unc, nch, nte, ted, eed, lea, eas, ase, nic, ice, sha, hal, all, cin, ine, nem, ema, wat, atc, tch, fil, ewa, hel, whr, sds, you, lik, ike, com, ome hen, giv, ive, cal, now, int, nto, tow, own, tod, oda, day, ilm, hav, ave, cou, oup, upl, dei, rin, ink, nks, aft, fte, erw, fou.	Average word length, Average sentence length, FW/CW ratio, Simpson index, Yule K, average contraction, average characters, type/token ratio

It should be noted that, at an early stage of research, analytical methods were tried with the non-adjusted and non-dimensionality reduced data matrices, but, the results were almost indistinguishable from those obtained with length-adjusted and dimensionality-data matrices. And this is because the created corpus was initially balanced for size and the unnecessary or unimportant variables were removed from it. However, the length-adjusted and dimensionality-reduced data matrices were used here for unbiased cluster analysis.

5. Method and Analysis

Thus far, each stylometric identifier was computed for its value/frequency in the 221 text messages. Attribution analysis was made on the basis of which of known-author text messages had the greater number of identifiers closer in similarity to the questioned text(s). Originally, the counting of stylometric identifiers was on a strictly numerical basis, *i.e.* if one participant had 5 identifiers and the other had 3 identifiers then the participant with the greater number of identifiers closer or similar to the respective identifiers in the anonymous text message was deemed to be the more probable sender. Given the data was too large, in terms of the number of variables and of the number of text messages used, it was difficult for the attribution analysis to be readily interpretable by direct inspection. To see the similarities and differences, I needed help, and that was what cluster analysis provided.

5.1 Cluster Analysis

In this test, cluster analysis technique was used to identify the relative degrees of similarity among the 221 text messages on the basis of their respective stylometric identifier values and represents the similarity pattern of the text messages in an intuitively interpretable graphical format. Cluster analysis is an exploratory multivariate data analysis technique for examining objects and grouping the objects into patterns or clusters based on some proximity (similarity/dissimilarity) measure. Cluster analysis is not a single method, but a family of related methods. Within each type of methods a variety of different specific methods and algorithms exist. In general, the working principal of cluster analysis is that objects in the same cluster have a small proximity from one another, while objects in different clusters are at a large proximity from one another. Similar objects should appear in the same cluster, and dissimilar objects in different clusters. (Everitt et al. 2001).

Given that hierarchical clustering provides more information than non-hierarchical ones, my only choice was to use hierarchical clustering on a 221 x 25 D_{POS}, 221 x 100 D_{FW}; 221 x 100 D_{bigram}, 221 x 100 $D_{trigram}$, and finally a 221 x 7 D_{tokens} to classify to identify the main clusters and also identify their constituency relations relative to one another as well as their internal structures. However, several hierarchical clustering methods were tried and we selected the method that gave the most intuitively clearest results about the constituency structure of the 221 text row vectors. In the current case, this was Beta-Flexible Clustering: the distance between two particular clusters CB and CC is the smallest of all distances, the flexible beta starts by merging them to form cluster CD. A user supplied/specified parameter β is then used to calculate the distance DAD, between CA and CD, which is defined as follows:

$$D_{AD} = (D_{AB} + D_{AC}) \times (1- \beta)/2 + D_{BC} \times \beta \quad (6)$$

where $-1 \leq \beta \leq 0$.

The process is repeated until all clusters belong to one single cluster. In this test, values for parameter β were chosen as -0.25, as recommended by Lance, G.N. and Williams, W.T. (1967) and Milligan, G. W. and Martha, C. Cooper (1987).

5.2 Attribution Analysis

The attributional analysis was based on all of the stylometric identifiers selected to describe the 221 text messages. Flexible Beta clustering tested the 221 text messages in pairs of known text messages with the questioned ones (i.e. the 10 text messages of interest)**.** For each participant pair of text messages of interest a text messages was chosen as the questioned text as described above. This text message was then compared with one each of the known participant pairs and with the other text messages in the created corpus. Flexible Beta clustering computed the similarity correlation coefficient between all pairs of texts and stored all the similarity values in a similarity/dissimilarity matrix. In the current case of Flexible Beta clustering, the similarities between all text messages were measured through Correlation Coefficient (Product-moment correlation). The similarity between two text message profiles was calculated as the correlation between the two profiles taken on by the two text vectors. This is expressed by the function:

$$S_{i.j} = \frac{\sum_{k=1}^{N}(C_{k,i} - \overline{C}_i)(C_{k.j} - \overline{C}_j)}{\sqrt{\sum_{k=1}^{N}(C_{k,i} - \overline{C}_i)^2 \sum_{k=1}^{N}(C_{k,j} - \overline{C}_j)^2}} \quad (7)$$

The reason for using Correlation Coefficient is that the aim here is to measure the similarity in patterns across POS usage profiles, function word usage profiles, word-bigram usage profiles, character trigram usage profiles, and token-based usage profiles regardless of overall magnitude and this measure is not influenced by differences in scales between data row vectors. It is expected that correlation between text messages using Flexible Beta clustering will not in general be affected by some text messages having larger average values or variations in their values. Thus, two text message vectors are perfectly similar when simply they have the same profiles regardless of overall magnitude. That is, two test row vectors have the same correlation coefficient (r=1), which implies to have a same pattern, but the distances are not equal.

5.3 Results

In the case of the five data matrices (POSs, FWs, Word bigrams, Character tri-grams, and Token based identifiers), consisting of 221 text messages by 211 participants there were possible 24310 repeated steps of merging clusters (to link all possible paired clusters) and were 110 participant pair similarity/dissimilarity comparisons to be made.

- Parts of speech dataset

The 211 known text messages and the 10 anonymous text messages in this corpus were described by 21 parts of speech. For this corpus a neutral result was obtained; five anonymous text messages were correctly attributed (person 1 P202, P203; person 4 P205, P211; person 5 P209, P214; person 9 P210, 204; person 6 P207, 212) and the five other anonymous text messages were not correctly attributed (person2 p45, p216; person 3 p179, p217; person 7 p178, p219; person 8 p208, 213, and person 10 P206, P220). Thus the attribution analysis accuracy rate for the data matrix of parts of speech (the number of text messages correctly attributed) was at 50%. While the syntactic identifiers have proven effective in certain authorship attribution and identification situations, they have not been sufficiently good at handling the subtlety of general short-text matching. Short texts often represent rich content of different syntactic identifiers, their relations are also complicated, and more sophisticated patterns are required for comparing and matching the two short texts. The result of POS analysis is presented in Figure (4), which is zoomed-in for clarity due to large size of the dataset.

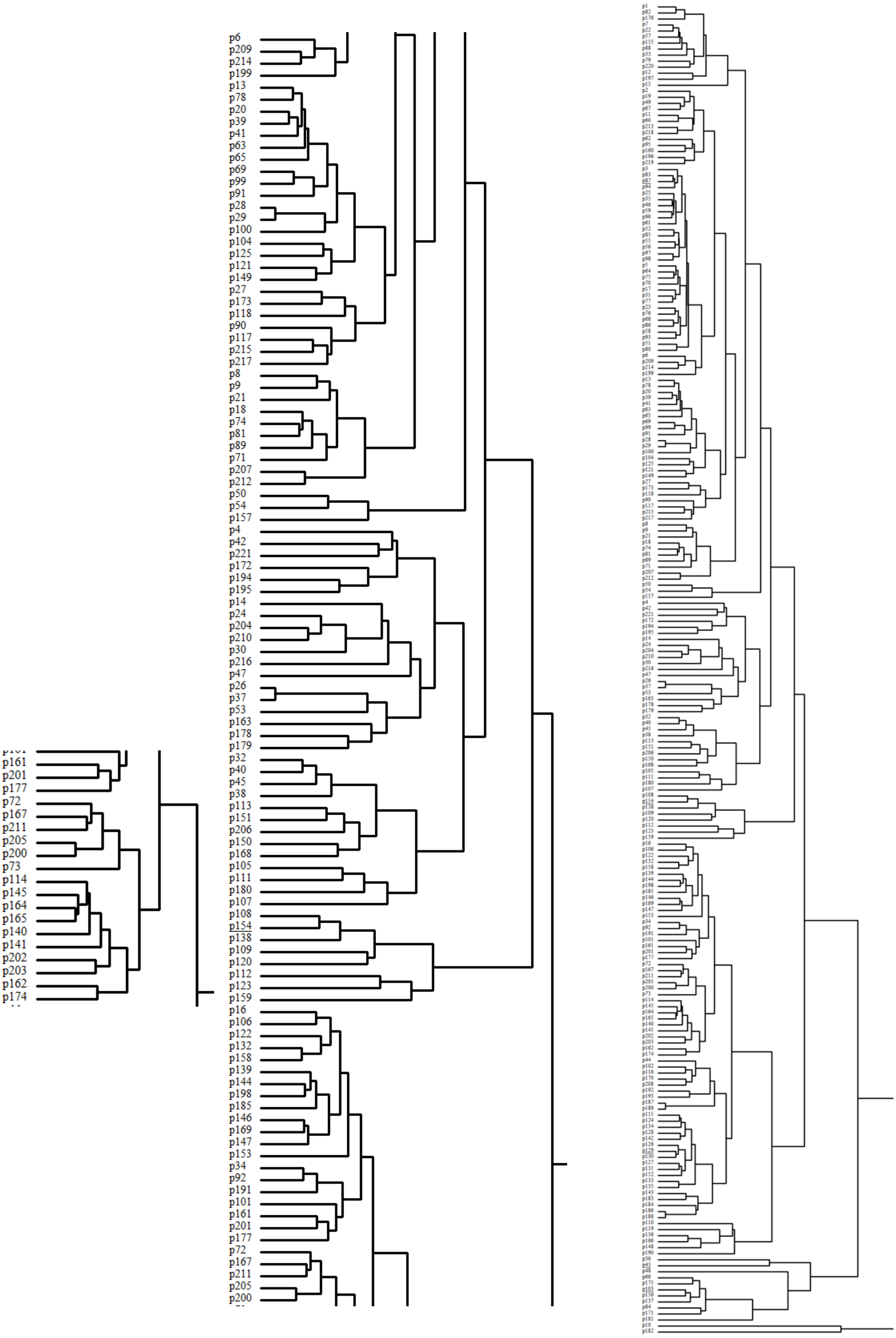


Figure 4. POS analysis

- FWs dataset

In this dataset 6 of the 10 anonymous text messages were correctly attributed, yielding an accuracy rate of 60%. For example, in the case of the analysis of person number 9 there were 4 text messages (p210, p202, p203, p204) to examine: each text message examined successfully, and the same is true for person number 6 (p207 and p212), person number 10, (p206 and p220), and person number 5 (p209 and p214)–but in the case of person 8 (p208 and p213), for example, or person number 1 (p40 and p218) the text messages weren't attributed correctly. Thus, all in all, the attribution test was successful between those six participants. As stated above, six anonymous text messages were correctly attributed due to their low variability usages within a participant's own known and disputed text messages and high variability usages across all participants in the entire dataset. The results are presented in Figure (5), which is zoomed-in for clarity due to large size of the dataset.

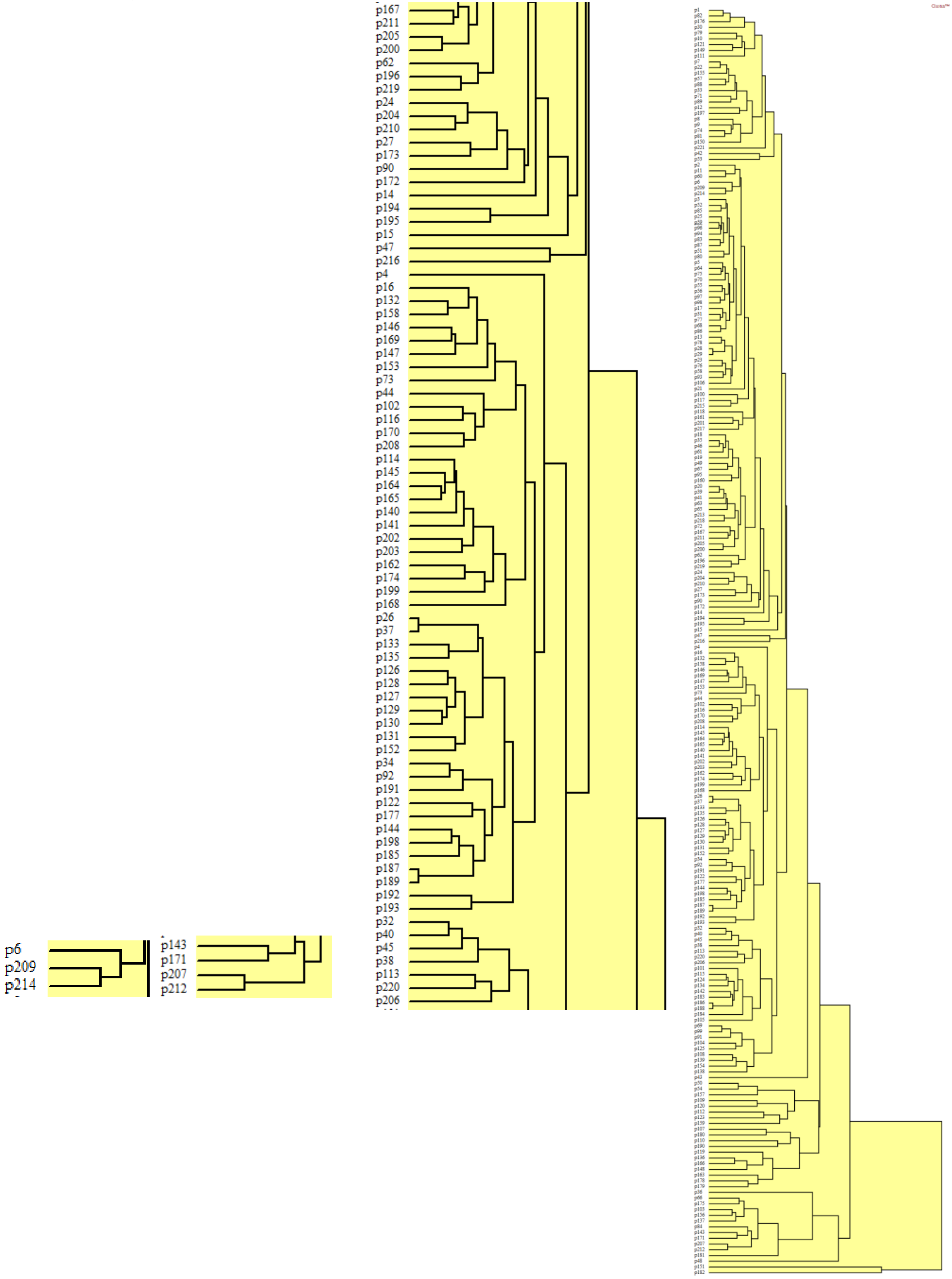


Figure 5. FWs analysis

- Word bi-grams dataset

This dataset consisted of 100 word bigrams each from the 211 known text messages and the 10 anonymous text messages. For this dataset a poor result was obtained since the analytical wasn't able to capture the more significant word bi-grams used by each participant. Four anonymous text messages were correctly attributed (person 9 p204, p210; person 6 p207, p212; person 5 p 209, p214; person 4 p205, p211) and five anonymous text messages were not correctly attributed (person1 p40, p218; person 2 p45, p216; person 3 p179, p217; person 7 p178, p219; and person 8 p208, 213; person 10 p206, p220). Thus the attribution analysis accuracy rate for the dataset of word bigrams (the number of text messages correctly attributed) was at 40%, and this is because, as was experimentally found, not every word combinations were encountered in a given text message because the representation produced by these identifiers was very sparse to account for all the possible combinations between words and this made it difficult to be captured by a clustering method used. The result is presented in Figure (6), which is zoomed-in for clarity due to large size of the dataset.

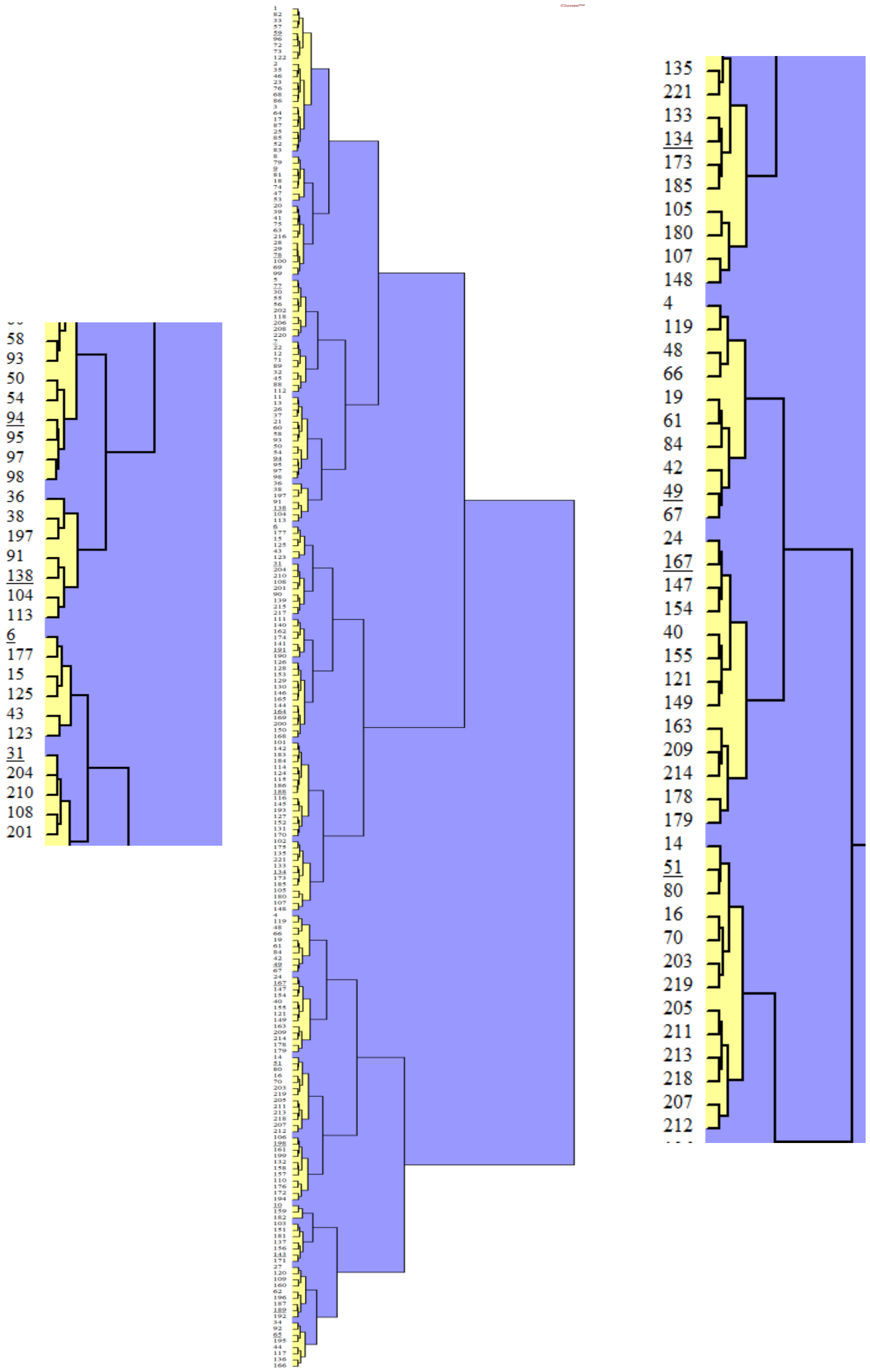


Figure 6. word bi-grams analysis

- Character triple grams dataset

This dataset consisted of 100 character triple grams. In this analysis, only four anonymous text messages were attributed correctly (person 5 p209, 2014; person 6 p207, p212; person 4 p211, p205; person 9 p202, p203). Accuracy rate of the character tri grams couldn't reach more than 40%, likely due to the small optimal number *n* or of variable *n* used in the test given that it was developed to capture only character 3-gram strings, representing sub-word or syllable like data. The result is presented in Figure (7), which is zoomed-in for clarity due to large size of the dataset.

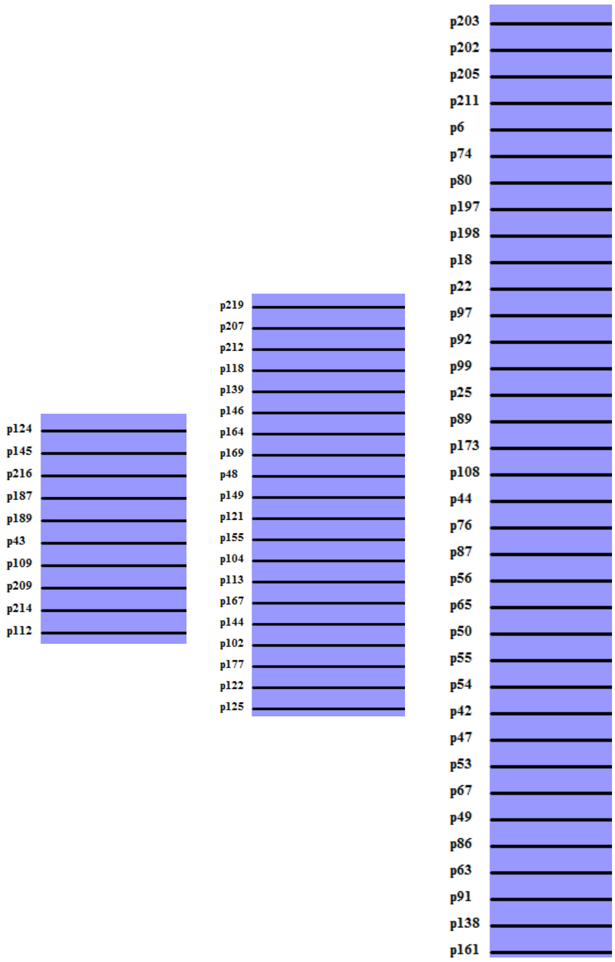


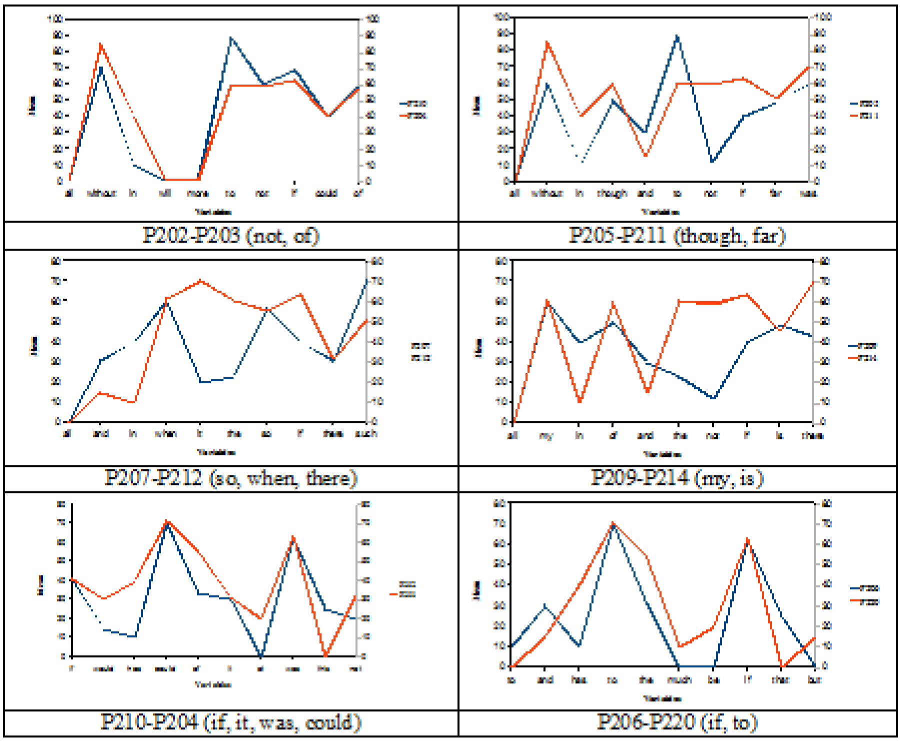
Figure 7. Character tri-grams analysis

- Token based features dataset

This dataset consisted of 8 token-based identifiers; identifiers such as the length of words and sentences as well as the density of function to content words and the richness of the vocabulary. The analysis of token-based identifiers gave the worst performance, where the results of this analysis were not considered here. This shouldn't come as a surprise since the text message genre and the dataset sample size greatly influenced the values measured in the measurement stage.

To sum up, the attribution results showed that the accuracy rate was at 60 % for the function words data matrix since only 6 of the 10 anonymous text messages were correctly attributed, while the accuracy rate for the parts of speech tag frequencies data matrix stood at 50% since only 5 anonymous text messages were correctly attributed. The attribution results also showed the accuracy rate for both word bigrams data matrix and character tri-grams data matrix was equally 40% since only 4 anonymous text messages were correctly attributed, while the accuracy rate for the token-based identifiers demonstrated that these identifiers aren't sufficient for the task of authorship attribution and identification of short text messages since all anonymous text messages were not correctly attributed.

Text message codes for all the texts of interest with centroid based identifiers for FWs and POS datasets are shown in Table (4), and their centroid analyses are shown in Figure (9) and Figure (10), respectively:



P202-P203 (not, of)

P205-P211 (though, far)

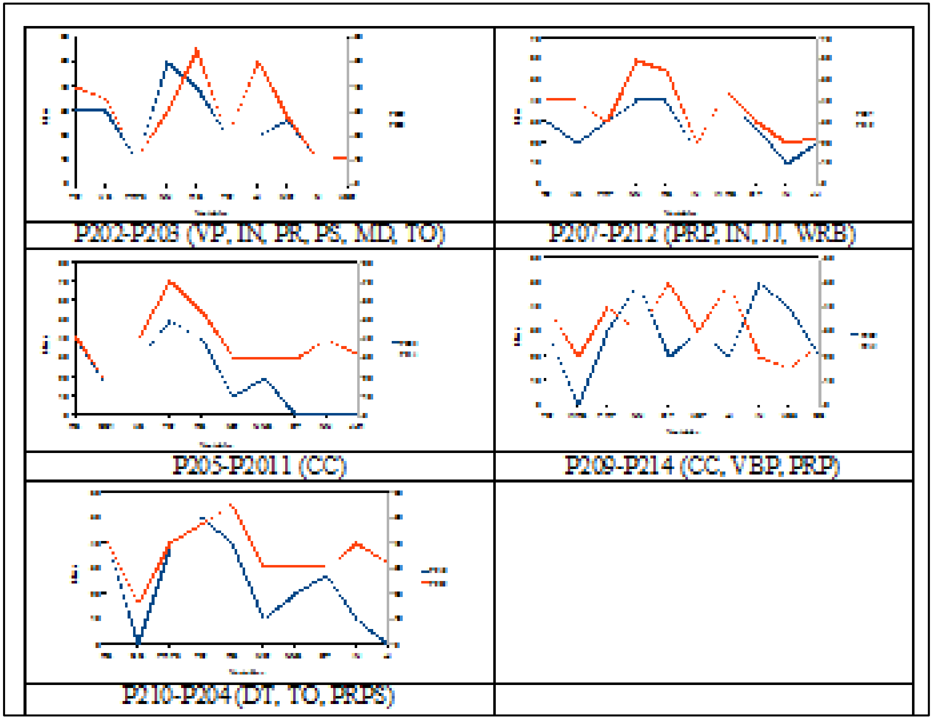
P207-P212 (so, when, there)

P209-P214 (my, is)

P210-P204 (if, it, was, could)

P206-P220 (if, to)

Figure 8. FWs centroid based analysis



P202-P203 (VP, IN, PR, PS, MD, TO)

P207-P212 (PRP, IN, JJ, WRB)

P205-P2011 (CC)

P209-P214 (CC, VBP, PRP)

P210-P204 (DT, TO, PRPS)

Figure 9. POS centroid based analysis

5.4 Testing of Results

The results presented may lead to an obvious question: Could the 60% result have obtained merely by chance? As one might argue that if a cluster analysis is applied on any given number of texts, a subset of clusters can always be resulted, even if there is no actual grouping of the texts. My answer simply is that this is an usual result and doesn't arise merely by chance. If the cluster P of X any cluster of interest wasn't similar to that of Y it would have been placed in distant cluster of its own. To support this answer, I compared the resulting model with another model that examines the same dataset: a clear convergence on one particular cluster model was held to support the validity of that model with respect to the data. Two hierarchical linear methods (Complete and Flexible Beta clusterings) applying Squared Euclidean Distance and one non-linear method (SOM) were used here. Specifically: SOM is a nonlinear method based on preservation of data topology; Complete and Flexible Beta clustering are both linear methods based on preservation of distance relations in data space, though they differ in how distance among clusters is defined.

Only the validation results obtained from the analysis of function words was shown because it is above chance level. The results from POS, word bigrams, character tri grams are not encouraging and therefore not validated.

5.4.1 Hierarchical Complete Clustering

In this method, the distance between two clusters A and B is based on the data vectors in each cluster that are furthest apart or furthest neighbors (longest distance)

5.4.2 Hierarchical Flexible Beta Clustering

See above.

5.4.2.1 Squared Euclidean Distance

In the current cases of Complete and Flexible Beta clusterings, the proximity between row vectors was measured through Squared Euclidean Distance. The reason of applying this measure is that it Squared Euclidean Distance is significantly affected by differences in scale. Two data objects will be judged to be different if they have differing overall incidences even when they follow a common pattern. The proximity between two vector profiles is calculated as the Squared Euclidean Distance between the two profiles taken on by the two vectors. Euclidean distance is the actual geometric distance between vectors in the space and Euclidean distance is the square root of the sum of the squared differences in the variables' values. This is expressed by the function:

$$d_{Euclid(BC)=}\sqrt{(X_{B-}X_C)^{2}+(Y_B-Y_C)^{2}} \tag{8}$$

Specifically: SOM is a nonlinear method based on preservation of data topology; Complete and, Flexible Beta clustering are both linear methods based on preservation of distance relations in data space, though they differ in how distance among clusters is defined.

The clustering results are presented in Figure (10) Complete and in Figure (11) Flexible, respectively.

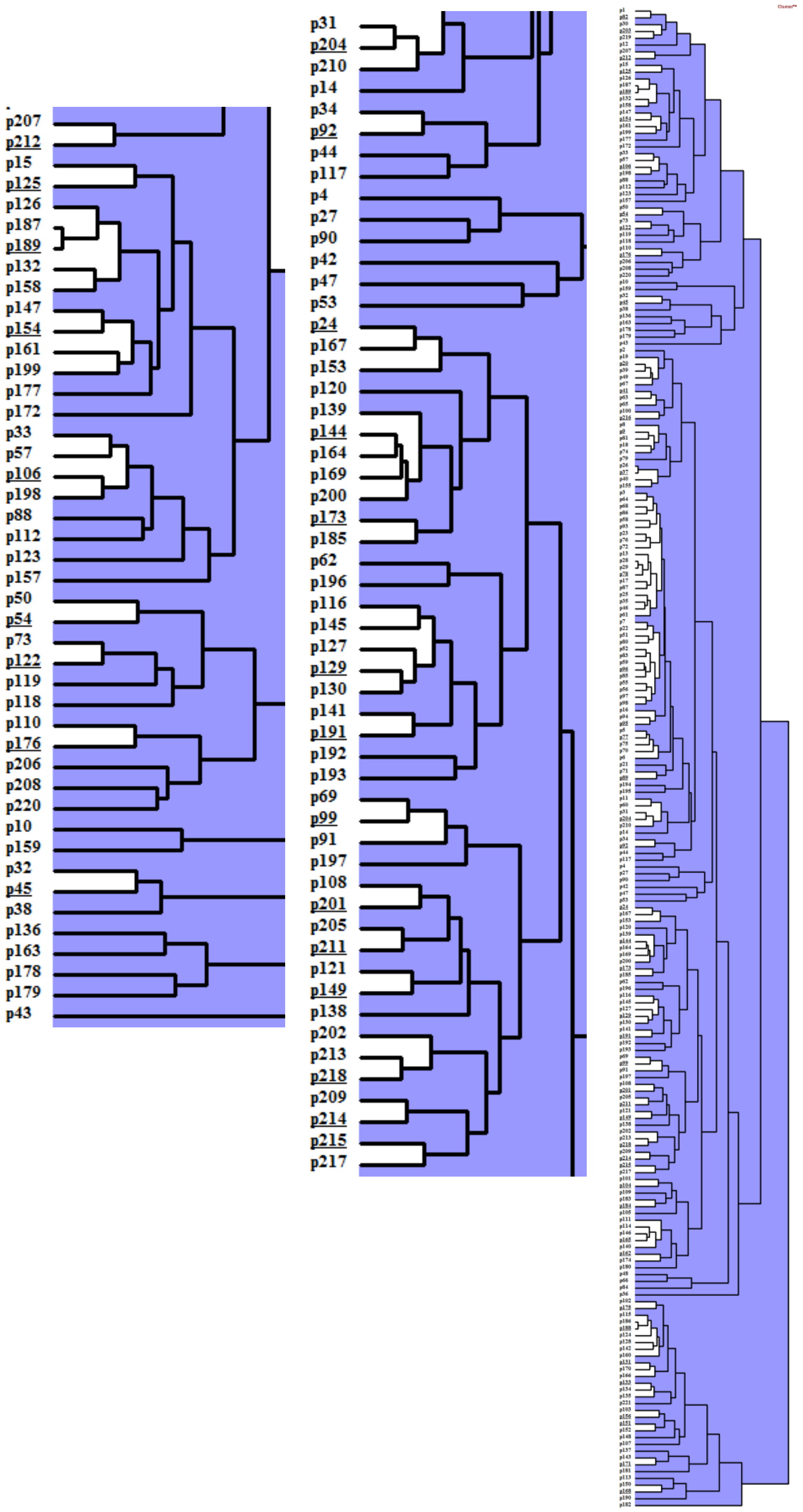


Figure 10. Testing analysis complete

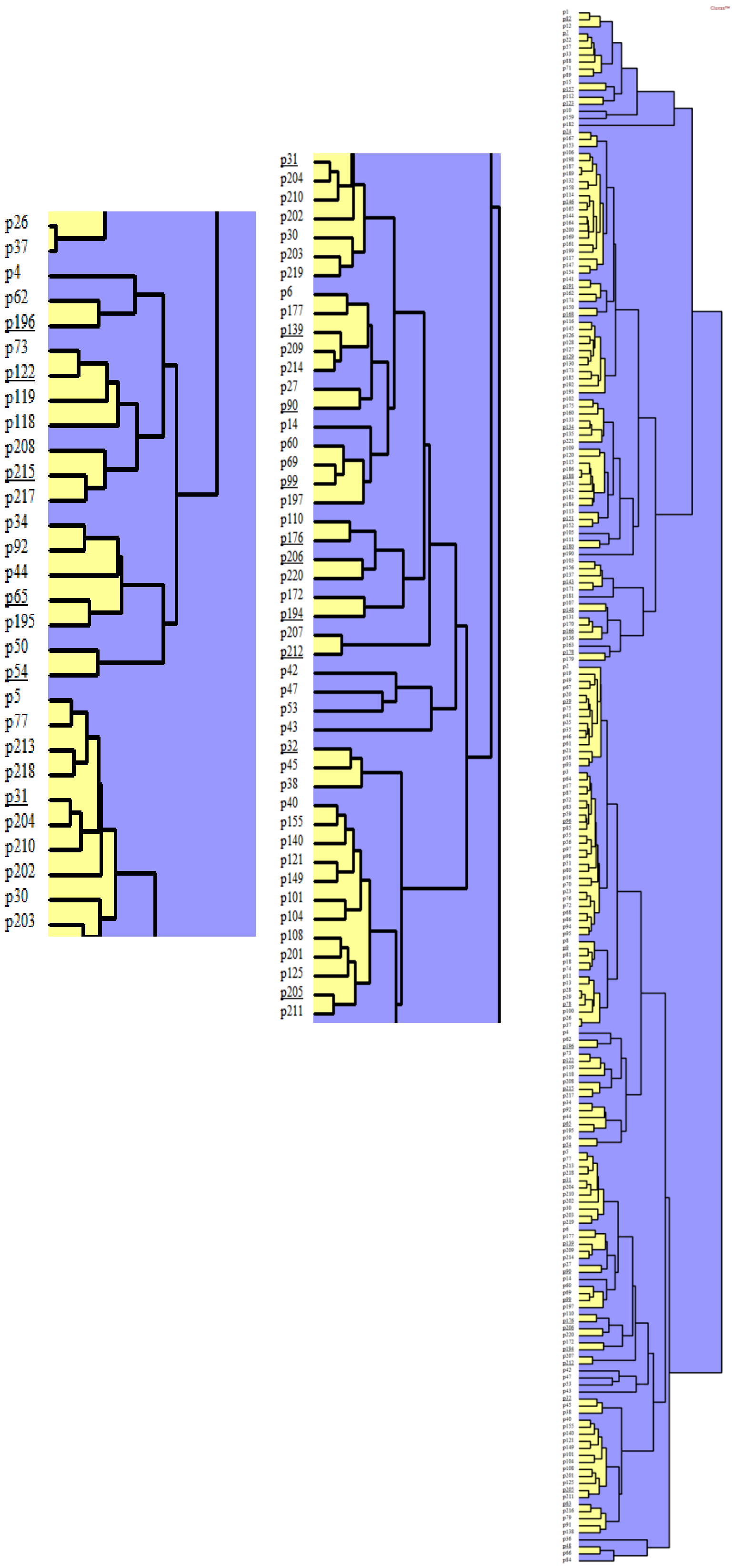


Figure 11. Testing analysis Flexible

5.4.2 Self-Organizing Map (SOM) U-Matrix

The unified distance matrix or U-matrix is a representation of SOM that calculates the nonlinear distances between data vectors and is presented with different colorings (Kohonen, 2001). It is based on preservation of data topology. SOM U-matrix generates graphical representations in two-dimensional space such that, given a suitable measure of proximity, vectors which are spatially or topologically relatively close to one another in high-dimensional space are spatially or topologically close to one another in their two dimensional representation, and vectors which are relatively far from one another in high–dimensional space are clearly separated, either by relative spatial distance or by some other graphical means, resulting—in the case of nonrandom data—in a configuration of well-defined clusters. The analysis was a two-stage process. The first was the training of SOM by loading all the vectors comprising DFWs into the input space. The second was the generation of the two-dimensional representation of the DFWs on the map. For each vector, the values in the input space were propagated through all the connections to the units in the lattice. Because of the variation in connection strength, a given vector activated one unit more strongly than any of the others, thereby associating each vector with a specific unit in the lattice. When all the vectors had been projected in this way, the result was a pattern of activation across the lattice. The U-matrix representation of SOM output used the relative distance between connection vectors to find cluster boundaries. Specifically, given a 221 × 100 output map DFW, the Euclidean distances between the connection vector associated with each map unit and the connection vectors of the immediately adjacent units were calculated and summed, and the result for each was stored in a new matrix UDFW, having the same dimensions as DFW. U was plotted using a color coding scheme to represent the relative magnitudes of the values in UDFW in which a dark coloring between the vectors corresponds to a large distance and, thus, represents a gap between the values in the input space. A light coloring is the boundaries between clusters or the vectors, indicating that the vectors are close to each other in the input space. Light areas represent clusters and dark areas cluster separators. Any significant cluster boundaries will be visible. Since the overall cluster structure is known from the immediately preceding section, some text messages that are not in the immediate neighborhood of the clusters of interest are deleted from this analysis, given in Figure (12), to avoid crowded plots or overloading and thereby obscuring the cluster results.

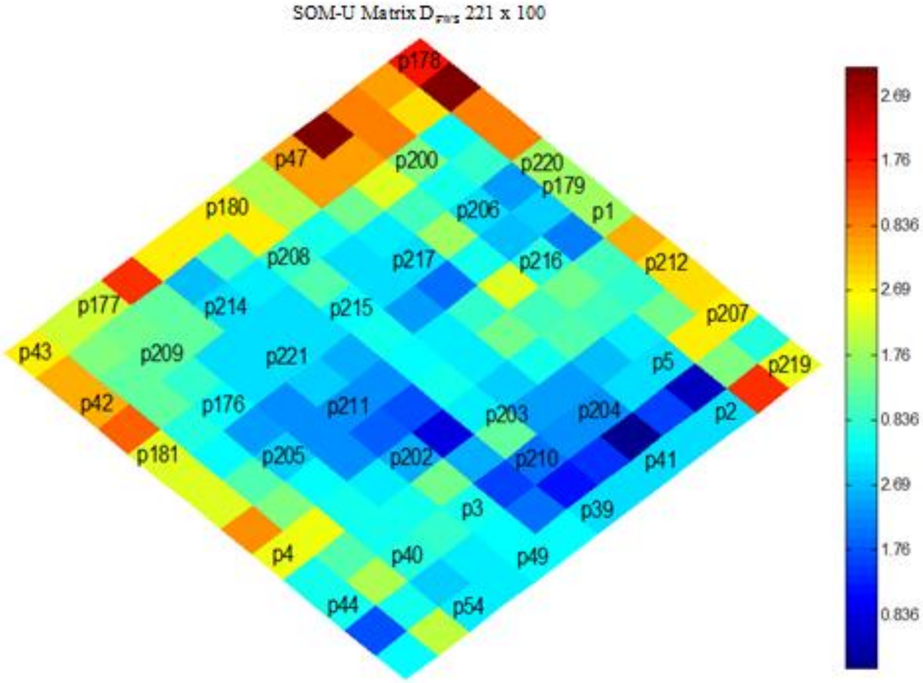


Figure 12. SOM testing analysis

The validation result presented in Figure (12) is shown to be identical to the experimental results obtained by the Flexible Beta clustering through correlation coefficients. All pairs of data points are grouped into clusters based on their distance values. For 60 % of the anonymous text messages the attribution analysis identified the author correctly: person number 9 (p210, 204), person number 6 (p207 and p212), person number 10, (p206 and p220), and person number 5 (p209 and p214), person number one (p202, p203), and person number 4 (p205, 211). This result provides a justification for the validity for the experimental results and for the proposed assumptions made to derive the clustering model.

6. Summary and Conclusions

The test in this study was conducted to provide a platform and support for the anonymous short text messages identification. More specifically, the aim was to determine, on the one hand, whether it is possible to identify the sender of a very short anonymous text message based on the style identifiers selected and to determine which types of identifiers make it possible to do so on the other. For this aim, I designed a simple corpus of 221 single-topic and real-life text messages and applied different Stylometric analytical techniques, I processed the text messages, and, from these, I generated five main datasets: parts of speech, function words, word bigrams, character tri grams, and token based identifiers. From each dataset, I examined and compared the text messages to 221 participants, and from each of these text messages, I selected the most important set of style identifiers that represented the linguistic profile of the 221 participants. I know empirically that if two linguistic profiles of real life text message have a similar profile representation (as measured by correlation coefficient) then they are likely to be similar. Four different types of hierarchical clustering modellings were used to execute the clustering task with each of these linguistic profiles. The identification analysis showed different attribution results. This means that each attribution analysis provides some information about the sender of the anonymous test messages. Function word usages showed the best attribution results, with a success rate of only 60 %. Therefore function words are indicative of the sender of anonymous short text messages based on 100 function words. With the parts of speech only 50% success rate was achieved based on 21 parts of speech. For all other identifiers an attribution result below this level was achieved: both word bi grams and character tri grams achieved 40%, and token-based identifiers achieved the worst result. This result was tested by using three other different clustering methods and I compared the test results to that of original result and found no significant changes between the two.

While this result is not what I hoped for, it remains the best outcome a test can achieve under the available data: extreme conditions, such as large number of authors, short text message sizes (about 2-3 lines /35-40 words in length), and single-genre/topic data sets. This tested result is within the expectation line and even considerably better if larger text message sizes (about 5 to 6 lines/100 words) per participant were available to the analytical methods to be analyzed. My conclusion is that there still is a need for more thorough testing with an expanded profile size that contains more than 100 words long for each participant that will identify a suspect person in very short text messages with higher than the success rates that obtained in the current application.

References

Aljumily, R. (2015). Hierarchical and Non-Hierarchical Linear and Non-Linear Clustering Methods to “Shakespeare Authorship Question”. *Social Sciences*, *4*, 758–799. https://doi.org/10.3390/socsci4030758

Aljumily, R. (2017). *Identifying Anonymous Online Message Senders: A Proposal Toward a Linguistic Fingerprint Biometric Database (LFBD)*. https://doi.org/10.2139/ssrn.3093279

Argamon, S., & Levitan, S. (2005). *Measuring the usefulness of function words for authorship attribution*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.6935&rep=rep1&typ e=pdf.

Belew, R. (2000). Finding out about: *A cognitive perspective in search engine technology and the WWW*. Cambridge: Cambridge University Press.

Church, K., & Gale, W. (1995a). Poisson mixtures. *Natural Language Engineering*, *1*, 163-190. https://doi.org/10.1017/S1351324900000139

Church, K., & Gale, W. (1995b). Inverse Document Frequency (IDF): A Measure of Deviation from Poisson. *In Proceedings of the Third Workshop on Very Large Corpora*, 121-130.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster Analysis* (4th ed.). London: Arnold.

Grieve, J. W. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and linguistic Computing*, *22*, 251-270. https://doi.org/10.1093/llc/fqm020

Holmes, D. I. (1994). Authorship Attribution. *Computers and the Humanities, 28*(2), 87-106. https://doi.org/10.1007/BF01830689

Holmes, D. I. (1998). Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing, 13*(3), 111-117. https://doi.org/10.1093/llc/13.3.111

Holmes, D. I., & Kardos, J. (2003). Who Was the Author? An Introduction to Stylometry. *Journal Chance*, *16*(2).

Houvardas, J., & Stamatatos, E. (2006). N-gram feature selection for authorship identification. *Artificial Intelligence*, *4183*, 77-86.

Kohonen, T. (2001). *Self-Organizing Maps* (3rd ed.). Berlin: Springer. https://doi.org/10.1007/978-3-642-56927-2

Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies I. hierarchical systems.

Computer Journal, *9*, 373-380. https://doi.org/10.1093/comjnl/9.4.373

Milligan, G. W., & Martha, C. C. (1987). *Methodology Review: Clustering Methods*. https://doi.org/10.1177/014662168701100401

Milton, J., & Arnold, J. (2003). *Introduction to probability and statistics: principles and applications for engineering and the computing sciences* (4th ed.). McGraw-Hill.

Pyle, D. (1999). *Data preparation for data mining*. San Francisco, CA: Morgan Kaufmann Publishers.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Sp ärck, J. K. (1974). Automatic indexing. *Journal of Documentation*, *30*(4), 393-432. https://doi.org/10.1108/eb026588

Sp ärck, J. K. (1972). Exhaustivity and specificity. *Journal of Documentation*, *28*, 11-21.

Sp ärck, J. K. (2004). IDF term weighting and IR research lessons. *Journal of Documentation*, *60*, 521-523. https://doi.org/10.1108/00220410410560591

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American society for information science and technology*, *60*, 538-556. https://doi.org/10.1002/asi.21001

van Rijsbergen, C. (1979). *Information retrieval* (2nd ed.). London: Butterworths

Copyrights