# Hello & Goodbye: Conversation Boundary Identification Using Text Classification

**Conference Paper** · June 2018

DOI: 10.1109/ISSC.2018.8585376

Jonathan Dunne
National University of Ireland, Maynooth
15 PUBLICATIONS 74 CITATIONS

SEE PROFILE

David Malone
National University of Ireland, Maynooth
143 PUBLICATIONS 2,923 CITATIONS

SEE PROFILE

# Hello & Goodbye: Conversation Boundary Identification Using Text Classification

Jonathan Dunne
Hamilton Institute
Maynooth University
Email: jonathan.dunne.2015@mumail.ie

David Malone
Hamilton Institute
Maynooth University
Email: david.malone@mu.ie

***Abstract*—One of the main challenges in discourse analysis is the process of segmenting text into meaningful topic segments. While this problem has been studied over the past thirty years, previous topic segmentation studies ignore crucial elements of a conversation: an opening and closing remark. Our motivation to revisit this problem space is the rise of instant message usage. We consider the problem of topic segmentation as a machine learning classification one. Using both enterprise and open source datasets, we address the question as to whether a machine learning algorithm can be trained to identify salutations and valedictions within multi-party real-time chat conversations. Our results show that both Naïve Bayes (NB) and Support Vector Machine (SVM) algorithms provide a reasonable degree of precision(mean F1 score: 0.58).**

## I. Introduction

Real-time group chat applications are seen as a way to improve productivity within mobile teams [1]. By using traditional conversational techniques, discussions can take place irrespective of location or time-zone. Small to medium (SMEs) businesses cite multiple benefits of real-time chat rooms. Such advantages include, brainstorming, client conferencing, customer support and distance learning [2]. A recent survey by ReportLinker found that while e-mail is a primary source of communication, group messaging application use is on the rise [3].

However, there are a number of drawbacks to group chat applications. Often cited are the problems with continuous partial attention (i.e. routinely checking a conversation) [4] or the lack conversation summarisation [5]. It is the latter problem that can prove challenging for users, especially if they have been away from a group chat for a period of time. Businesses face challenges in this regard, as current group chat applications offer little or no chat summarisation functionality.

Text segmentation is a technique used to separate text into meaningful clusters. Such clusters may include sentences or topics. Previously, text segmentation research has focused on topic changes within written discourse. Such discourse comprises prose text [6]. However, in recent times attention has turned to conversational discourse such as real-time chat [7]. In an age of big data, coupled with the fact that businesses are using collaboration applications more than before [8], being able to segment chat conversations by topic may prove useful in the domain of information retrieval.

In this paper, we propose a technique that text segmentation practitioners can use to annotate conversations with an opening (salutation) and closing (valediction) remark. The core idea of this study is to demonstrate that by manually annotating conversation boundaries, a trained machine learning classifier algorithm can identify conversation boundaries using salutations and valedictions as a conversation perimeter. For topic modelling practitioners, identification of conversation boundary markers may improve text mining outcomes on a per-conversation basis.

This study contains research conducted on two real-time chat discourse datasets. Our first dataset is an enterprise dataset from a real-time collaboration application; our second dataset is an open source data set from an Internet chat relay (IRC) channel. We investigate a) the high frequency words and key collocations that are present in salutation & valediction messages and b) whether two specific machine learning classification algorithms can identify salutations, valedictions and conversation body text within multi-party chat discourse. The results of our study can be used to further the body research in the field of text segmentation.

The rest of the paper is structured in five sections: Section II gives some description of study background and related works. Section III describes the enterprise dataset. Section IV provides analysis and methodology. It is followed by section V that discusses the result. Finally, the conclusion and future work are described in section VI.

## II. Background and related research

### A. Text Classification

Text classification is a subset of document classification, whereby text is required to be labelled as a specific class or category. Classes are selected from an established hierarchy of existing classes. For example, text may be classified by subject, author or emotive tone.

The classification task was traditionally a manual one. However, in recent times, due to the advent of large corpora of text data and relatively cheap computing power, the task is mainly conducted using a machine learning algorithm with varying degrees of success [9].

Today text classification by computers is used to solve many concrete problems such as sentiment detection (i.e. detecting

978-1-5386-6046-/18/$31.00 © 2018 European Union

positive or negative film reviews), e-mail sorting (i.e. sort e-mails sent by family, business colleagues or a spambot).

### B. Machine Learning

Machine learning is an area of computer science that allows computers to learn the outcome of a task without being explicitly being programmed to do so [10]. Machine learning begins by observing data directly and using this knowledge to infer patterns in data and make decisions or predictions on additional examples. The phrase "Machine Learning" was first coined by Arthur Samuel in 1959 while working at IBM [11].

Machine learning can employ different algorithms to provide output. These algorithms can be categorised as supervised, unsupervised, semi-supervised and reinforcement.

Supervised learning is probably the most common type of algorithm used today. The core idea that prior labelled data (known as training data) is used to generate a corresponding matching output from another set of data (known as test data). Ideally, the machine learning algorithm can generalise the training data in a meaningful way to determine a classified label from unseen data [12]. Examples of supervised algorithms include NB, SVM, decision trees and random forest.

With unsupervised learning, prior data is neither labelled or classified. In this case, an algorithm is used to cluster data around data that is inferred as being similar. The main aim of unsupervised learning is to provide exploratory data analysis by inferring hidden patterns or groups [13]. Examples of unsupervised algorithms are k-means clustering, Gaussian mixture models and hidden Markov models.

Semi-supervised learning is a hybrid of both algorithms, typically a small amount of labelled data is used to cluster data into a set of known groups. Reinforcement learning is an approach whereby an algorithm interacts with an environment to determine a set of actions to maximise a reward. This method allows of a level of 'ideal behaviour' to be inferred [14].

In the following subsections, we discuss two algorithms in more detail and summarise notable prior research in the field of text segmentation.

### C. Naïve Bayes

A NB classifier is a type of machine learning algorithm that uses Bayes theorem. This algorithm makes a strong (naïve) assumption of independence between each pair of features [15]. Despite the seemingly over-simplified assumptions of independence, NB has shown to be useful with solving real word problems most notably in the field of document classification and e-mail spam filtering [16].

### D. Support Vector Machine

SVM is an algorithm used in machine learning to solve classification and regression problems [17]. SVM represents labelled observations as points in space. As the points as plotted the algorithm determines what line best separates the labelled classes. This separation point is also known as a hyperplane. Ideally, a hyperplane with the largest distance between both sets of classes is preferable, as this makes it easier to distinguish between classes.

If a classification problem presents whereby a line is unable to separate the labelled classes successfully, SVM can use a non-linear classification. A "kernel trick" is used to perform a data transformation to create a high dimension feature space. As a result the hyperplane may be extended to a curve or a series of curves. The kernel trick was initially proposed as far back as 1964 by Mark Aizerman [18]. Vapnik et al. are credited with successfully incorporating the kernel trick to SVM in the early 1990's [19].

Due to SVM's flexibility, the algorithm has been used to solve many real world problems in the field of text and image classification (face recognition) [20]. Additionally in the field of bioinformatics, SVM is been shown to be an effective method to classify proteins [21].

### E. Studies Related to Text Segmentation

The purpose of text segmentation is to identify specific subregions of text within a corpus. The benefit of such a practice is to aid in the field of information retrieval, where topic boundary identification is a crucial problem. We discuss some of the leading contributions to the domain of topic boundary identification briefly.

One of the first studies (1991) in the field of text segmentation was conducted by Morris and Hirst [22]. The authors focused on the problem of lexical cohesion (chains of related words), by using a thesaurus as a knowledge base for computing lexical chains. Additional early contributors in the field of lexical cohesion include Kozima [23], who proposed a lexical cohesion profile, and Reynar [24] who outlined an improved method of locating discourse boundaries based on the previous method of lexical cohesion and a graphical technique call dotplotting.

Some years later, additional techniques have been used to tackle the problem of partitioning text into coherent segments. Beeferman et al. [25] introduce an exponential model to extract features that are correlated to the presence of boundaries. Their study used *Wall Street Journal* news articles and television news story transcripts. Galley et al. [26] propose a discourse segmentation technique using multi-party conversations. Their lexical cohesion algorithm demonstrated reasonable results when text extracted from the Brown corpus[1].

In more recent times (2012 onwards), new researchers used different techniques to research text segmentation. Nguyen et al. [27] proposed a Bayesian nonparametric model to discover the topics used in a conversation, topic shift and a person specific tendency to introduce new topics. The authors used transcripts from the 2008 presidential debate and a television programme called Crossfile. Brooks et al. [28] used a machine learning approach to identify effective state (e.g. joy excitement, confusion, frustration, anger and annoyance) on chat logs, using comprised of discussion from an astrophysics institute. Schmidt and Stone [29] use a combination of techniques

[1] http://clu.uni.no/icame/brown/bcm.html

TABLE I
SUMMARY OF DATASET METRICS AND BOUNDARIES

| Dataset | Enterprise | Ubuntudev-IRC |
|---|---|---|
| **Total # Messages** | 3261 | 4223 |
| **Total dataset duration (hours)** | 4822 | 86 |
| **Multi-line conversations annotated** | 257 | 207 |
| **Salutations** | 257 | 207 |
| **Valedictions** | 257 | 207 |

(i.e. Latent semantic analysis, text tiling, and pause detection) to detect topic changes in IRC chat logs, with limited success.

Rounding off our studies in this section, Uthus and Aha [30] surveyed research on the analysis of multi-participant chat. The authors conclude that chat data is difficult to analyse due to its unique characteristics due to the many problems the medium presents (e.g. Chat room feature processing, thread disentanglement, topic detection, summarisation and user profiling). This has caused many traditional text analysis techniques to prove unsuccessful. The authors suggest that given its prevalence of social communication, the domain represents an exciting research topic.

## III. Data sets

Text segmentation of social media/collaboration messaging can be a useful technique to improve the quality of text mining and summarisation tasks. By annotating conversation boundaries by their salutation and valedictions, we demonstrate how machine learning algorithms can be trained to identify such opening and closing remarks with a reasonable degree of precision.

The study presented in this paper examines four hundred and sixty-four manually annotated real-time chat conversations from two datasets. The details are summarised in Table I.

For each message, we noted whether it was a salutation (i.e. an opening remark, greeting etc.). Each subsequent message was read until a valediction (i.e. a closing remark, farewell or acknowledgement message) was found. With a conversation perimeter identified we assigned a numeric topic ID. A number of single-line topics were found as part of the annotation process. For this study, only multi-line topics (i.e. conversations with one distinct salutation and valediction) are considered as part of our analysis.

The first dataset analysed was from an enterprise instant message chat system which discussed cloud infrastructure problems. For our study, we reviewed approximately 3200 messages. As part of the review phase, we annotated 257 distinct conversations. The total time period analysed was approximately 4820 hours.

The second dataset[2]analysed was the open source Ubuntu dev IRC channel [31]. For our study, we reviewed approximately 4200 messages. As part of the review phase, we annotated 207 unique conversations. The total time period analysed was approximately 86 hours.

TABLE II
CONDENSED IRC CONVERSATION WITH CLASSIFICATION LABELS

| Date | User | Test | Classifier |
|---|---|---|---|
| 01/10/04 06:20 | <m_tthew> | fabbione: ahoy | salutation |
| 01/10/04 06:27 | <fabbione> | hey m_tthew | message-body |
| 01/10/04 07:04 | <mdz> | morning | message-body |
| 01/10/04 07:11 | <fabbione> | mdz: for the ati / flrdkjdjds driver... | message-body |
| 01/10/04 07:18 | <fabbione> | building now :-) | message-body |
| 01/10/04 07:18 | <fabbione> | brb | valediction |

This study aims to answer the following questions. First, what types of words and collations are contained within salutation and valediction messages? Second, can a classifier algorithm be trained to identify salutary and valedictory text from real-time chat messages, thus identify a conversation boundary?

### *A. Lexicography*

Lexicography is the study of vocabulary meaning and its use. In recent times, research has expanded to include a corpus based approach [32]. The benefit of a corpus based approach is as follows:

- What is the frequency of word usage?
- What is the frequency of word usage across multiple senses?
- Do words have a systematic association with other words?

The advantage of the corpus-linguistic method is that language researchers can analyse naturally occurring language text produced by a variety of authors to confirm or refute intuitions about specific language features using empirical data.

We aggregated both the salutation and valediction messages from each dataset into a single corpus. We then used the corpus linguistic tool #lancsbox [33] to analyse our salutation and valediction words. Our first research question asked a) what are the high-frequency words used and b) what interesting collations are present in salutation & valediction messages.

### *B. Chat boundary classification*

Our second research question asked, can a machine learning classifier algorithm be trained to identify text as a salutary and valedictory text from real-time chat messages with a reasonable degree of precision. For this research question, we constructed a multinomial classification experiment using three classes; salutation (opening message), message-body (neither an opening or closing message) and valediction (closing message). Table II provides an overview of the classification methodology.

There is an open question as to whether stop words (i.e. the most common words in a language) should be removed before classification. If stopwords are removed, it can remove "noise" from sentences and allow a classifier to focus on a subset of

[2] A copy of the annotated open-source dataset is available from the corresponding author upon request.

TABLE III
LIST OF THE MOST FREQUENT NON STOP WORDS

| # | Word | Frequency | # | Word | Frequency |
|---|---|---|---|---|---|
| 17 | thanks | 68 | 35 | jenkins | 36 |
| 20 | can | 64 | 36 | ok | 36 |
| 21 | now | 51 | 37 | update | 35 |
| 24 | will | 49 | 40 | get | 35 |
| 27 | today | 42 | 42 | please | 33 |
| 29 | not | 40 | 43 | need | 33 |
| 30 | as | 40 | 44 | morning | 33 |
| 31 | new | 38 | 45 | working | 31 |
| 34 | just | 36 | 48 | still | 29 |

text. However, the concern is that valuable text markers may be lost if such text is removed. We conduct our classification experiment on both the full text and with stop words removed.

Next, we conducted the following steps to train each algorithm and evaluate each training set using the python library scikit-learn[3].

- Each dataset was divided into a training, development and test set, in a ratio of 60% / 20% / 20%.
- Each training set was trained against both classifier algorithms.
- A development set was used to assess the performance of each algorithm.
- A test set was used to assess the performance of each algorithm to assess over/under-fitting.
- The above steps were repeated with the stop words removed from both datasets.

Note: For the SVM algorithm we assessed a total of thirty-six combinations of loss and penalty functions using the development set. The highest performing combination was then used to validate the training set, using the same parameters. The NB algorithm has no loss or penalty tuning parameters.

### C. Limitations of dataset

The dataset has some practical limitations, which are now discussed. The process of aggregating chat messages into a cohesive conversation is a subjective one. Every effort was made to assign messages to their most appropriate thread. We recognise that the process is subjective, and may be subject to type I errors.

The chat conversations that form part of this study are from an Ubuntu IRC developer channel and an Enterprise Cloud channel. While we hope these examples will be representative of technical discussion channels, it seems unlikely they will be typical of all types of channels.

## IV. RESULTS

### A. Lexicography

We tabulated a list of the fifty most common words found in both salutation and valediction text. Thirty-two of these words were found to be either stop words or usernames. We removed these words from our list. Table III provides a summary of the most common words used across the salutation and valediction corpus. We include the absolute rank of the word frequency in the overall corpus.

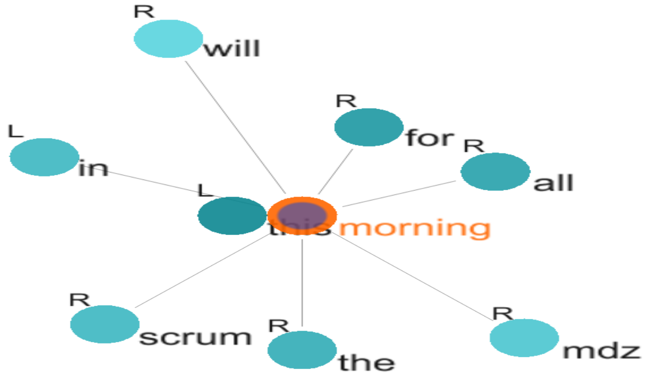

Fig. 1. Collocation plot for the word ‘morning’

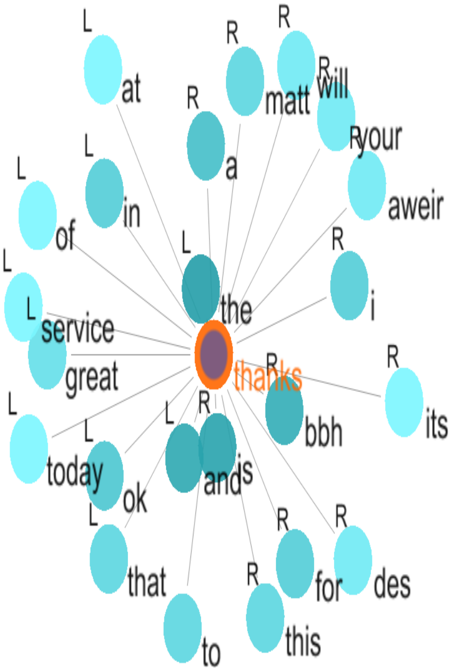

Fig. 2. Collocation plot for the word ‘thanks’

We choose two words: *morning* and *thanks* from Table III to explore collocations in more detail. We selected *thanks* as we observed that it was used quite frequently in a number of closing messages to acknowledge completion of some task. We selected *morning* as it appeared in a number of salutations as a form of greeting. Fig.1 & Fig. 2 illustrate a collocation graph for each word. A collocation graph is used to show how a given word is used in conjunction with other words, and

[3] http://scikit-learn.org/stable/

whether that word appears to the left or right of a collocated word.

### B. Chat boundary classification

Table IV provides a summary of the classification results. We note that highest mean precision, recall and F1 score was achieved with the Ubuntu IRC dataset with stop words removed. The lowest mean precision, recall and F1 score were against the combined Enterprise & Ubuntu IRC dataset with stop words removed. For SVM the Huber loss function with no penalty provided the highest mean scores.

## V. Discussion

### A. Lexicography

Our first research question asked a) what are the high-frequency words used and b) what interesting collocations are present in salutation & valediction messages. We note that thirty-two of the top fifty words were 'stop words', as these words are the most commonly used words in the English language this is unsurprising. Table III illustrates that *thanks, can* and *now* appear fifty times or more. We note that: *Thanks* (a noun or interjection used to express gratitude), *Can* (is an auxiliary verb used with a pronoun (e.g. you)) and *Now* (an adverb used to add a time dimension to an action or statement). Interestingly, neither a variation on the standard greeting or farewell (e.g. hello, goodbye) appeared in the top one hundred words. *Hi* was the 120th most frequent word while *later* was the 292nd most frequent word.

Looking at the collocation graph in Fig. 1 we see *morning* collocates with nine words. The most frequent collocates were *this* (left) and *for* (right). Interestingly, morning does not collocate with *good*, however it does collocate with a username *mdz* on six occasions. Fig. 2 shows the collocation graph for *thanks*. We can see that *thanks* collocates with twenty-nine distinct words. On twenty-four occasions, *thanks* has no collocates (i.e. used as a single word message), also on thirty-four occasions, *thanks* collocates with four distinct usernames (i.e. *bbh, matt, des, aweir*).

The key takeaway from this work is to demonstrate how complex and variable written discourse is. By adopting a corpus based approach, we can understand how language is used within a specific domain. The results of corpus analysis can be used define features for use as part of a deep learning architecture.

### B. Chat boundary classification

Our second research question asked, can a machine learning classifier algorithm be trained to identify text as a salutation or valediction from real-time chat messages and if so to what degree of precision. Looking at Table IV, a number of points are apparent. Overall SVM performed best in five of the six experiments, the highest mean precision, recall and F1 score achieved was with the Ubuntu dataset with stop words removed. The Huber loss function with no penalty provided the highest level precision across all experiments.

Combining the two datasets did not yield a significant improvement classifier performance. By combining the datasets, we added more variance to our training and test data, which degraded classifier performance. Interestingly, when stop words were removed from each dataset, we observed a slight increase in classifier performance except with the combined dataset. Our intuition suggests that by removing stop words, we removed some of the variance from both datasets.

Overall, the both classifier algorithms achived average or slightly better than average performance. We see two contributing factors. Firstly, we know there is variability in the language used within salutation and valediction text. We observed that a number of conversations start with an image, URL or emoji rather than a word. Secondly, for the message-body label, there is even more variability due to the many ways in which humans can express themselves in chat discourse.

The main benefit of such an experiment is to highlight the idea, that neither the NB nor SVM text classifier algorithms are suited to identify specific types of utterances in chat discourse with a high degree of precision. However, we acknowledge that increasing the level of training and test data would be reasonable for further experiments in this area.

## VI. Conclusion

The purpose of this study was to understand whether a text classification algorithm could be used to identify conversation boundaries using salutation and valediction text within a group chat context. Additionally, we adopted a corpus linguistic approach to identify lexical patterns within group chat conversations.

We found that both NB and SVM provide a modest level of precision identifying opening and closing remarks within group chat conversations. Additionally, we found that classifier performance varied little between datasets and that removal of stop words increased classifier performance on each data set. Combining datasets saw a slight decrease in classifier performance.

Furthermore, we found that a corpus-based approach can provide useful insights into the mechanics of opening and closing messages of a group chat conversation with the use of collocations.

These initial results may be of use to SMEs and researchers understanding the use of NB and SVM for identification of salutations and valedictions. By adopting a fine grained corpus-based approach, additional features may be developed to provide models with improved performance. We consider that a multi-feature classification model in the form of a neural network could be used to further our work.

In future work, we shall investigate the idea of whether multiple messages form a salutation and validation cluster. Additionally, we shall evaluate our datasets with two other classification algorithms: Decision trees and the ensemble method random forest using increased training and test data.

## References

[1] (2017) The Six Benefits of Real-Time Chat For Your Mobile Workforce. [Online]. Available: http://bit.ly/2GZ7Pk9

TABLE IV
SUMMARY OF TEXT CLASSIFICATION EXPERIMENTS

| **Dataset** | **Processing** | **Classifier** | **Label** | **Precision** | **Recall** | **F1** |
|---|---|---|---|---|---|---|
| Ubuntu IRC | None | NB | salutation | 0.52 | 0.55 | 0.54 |
| | | | message-body | 0.57 | 0.8 | 0.67 |
| | | | valediction | 0.55 | 0.3 | 0.39 |
| | | | **mean** | **0.55** | **0.55** | **0.53** |
| Ubuntu IRC | Tokenised & | SVM | salutation | 0.62 | 0.45 | 0.52 |
| | Stopwords removed | | message-body | 0.54 | 0.7 | 0.61 |
| | | | valediction | 0.62 | 0.6 | 0.61 |
| | | | **mean** | **0.59** | **0.58** | **0.58** |
| Enterprise | None | SVM | salutation | 0.55 | 0.62 | 0.58 |
| | | | message-body | 0.57 | 0.58 | 0.57 |
| | | | valediction | 0.56 | 0.48 | 0.52 |
| | | | **mean** | **0.56** | **0.56** | **0.56** |
| Enterprise | Tokenised & | SVM | salutation | 0.68 | 0.64 | 0.66 |
| | Stopwords removed | | message-body | 0.49 | 0.58 | 0.53 |
| | | | valediction | 0.57 | 0.5 | 0.53 |
| | | | **mean** | **0.58** | **0.57** | **0.57** |
| Combined | None | SVM | salutation | 0.59 | 0.52 | 0.55 |
| | | | message-body | 0.54 | 0.58 | 0.56 |
| | | | valediction | 0.52 | 0.53 | 0.52 |
| | | | **mean** | **0.55** | **0.54** | **0.54** |
| Combined | Tokenised & | SVM | salutation | 0.57 | 0.57 | 0.57 |
| | Stopwords removed | | message-body | 0.50 | 0.41 | 0.45 |
| | | | valediction | 0.49 | 0.58 | 0.53 |
| | | | **mean** | **0.52** | **0.52** | **0.52** |

[2] (2018) The Advantages of a Chat Room. [Online]. Available: http://bit.ly/2iy5qVS

[3] (2017) Can We Chat? Instant Messaging Apps Invade the Workplace. [Online]. Available: http://bit.ly/2Egv3ES

[4] (2017) Pros and Cons of corporate group chats. [Online]. Available: http://bit.ly/2Eda9m0

[5] (2017) Is group chat making you sweat? [Online]. Available: http://bit.ly/2nVcj2t

[6] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 9, no. 2-4, pp. 123–138, 2007.

[7] J. Weisz, "Segmentation and classification of online chats," 2008.

[8] (2017) Group-chat software sees explosive growth and intense competition. [Online]. Available: http://bit.ly/2nOuaJ8

[9] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[10] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development*, vol. 3, no. 3, pp. 210–229, 1959.

[11] F. Provost and R. Kohavi, "Glossary of terms," *Journal of Machine Learning*, vol. 30, no. 2-3, pp. 271–274, 1998.

[12] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2012.

[13] T. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised learning," in *The elements of statistical learning*. Springer, 2009, pp. 485–585.

[14] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[15] S. J. Russell, P. Norvig, J. F. Canny, J. M. Malik, and D. D. Edwards, *Artificial intelligence: a modern approach*. Prentice Hall Upper Saddle River, 2003, vol. 2, no. 9.

[16] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," in *Learning for Text Categorization: Papers from the 1998 workshop*, vol. 62, 1998, pp. 98–105.

[17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[18] M. A. Aizerman, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and remote control*, vol. 25, pp. 821–837, 1964.

[19] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

[20] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," in *Computer vision and pattern recognition, 1997. Proceedings., 1997 IEEE computer society conference on*. IEEE, 1997, pp. 130–136.

[21] L. Liao and W. S. Noble, "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships," *Journal of computational biology*, vol. 10, no. 6, pp. 857–868, 2003.

[22] J. Morris and G. Hirst, "Lexical cohesion computed by thesaural relations as an indicator of the structure of text," *Computational linguistics*, vol. 17, no. 1, pp. 21–48, 1991.

[23] H. Kozima, "Text segmentation based on similarity between words," in *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1993, pp. 286–288.

[24] J. C. Reynar, "An automatic method of finding topic boundaries," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 331–333.

[25] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Machine learning*, vol. 34, no. 1-3, pp. 177–210, 1999.

[26] M. Galley, K. R. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 2003.

[27] V.-A. Nguyen, J. Boyd-Graber, and P. Resnik, "SITS: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 78–87.

[28] M. Brooks, K. Kuksenok, M. K. Torkildson, D. Perry, J. J. Robinson, T. J. Scott, O. Anicello, A. Zukowski, P. Harris, and C. R. Aragon, "Statistical affect detection in collaborative chat," in *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2013, pp. 317–328.

[29] A. P. Schmidt and T. K. Stone, "Detection of topic change in irc chat logs," 2013.

[30] D. C. Uthus and D. W. Aha, "Multiparticipant chat analysis: A survey," *Artificial Intelligence*, vol. 199, pp. 106–121, 2013.

[31] (2017) Ubuntu irc logs. [Online]. Available: https://irclogs.ubuntu.com/

[32] D. Biber, S. Conrad, and R. Reppen, *Corpus linguistics: Investigating language structure and use*. Cambridge University Press, 1998.

[33] (2017) Lancsbox: Lancaster University corpus toolbox. [Online]. Available: http://bit.ly/2smurrz