Research Article

Akshay Mendhakar* and Darshan H S

Parts-of-Speech (PoS) Analysis and Classification of Various Text Genres

https://doi.org/10.1515/csh-2023-0002
Received April 12, 2023; accepted December 4, 2023; published online December 21, 2023

Abstract: Natural language processing (NLP) has made significant leaps over the past two decades due to the advancements in machine learning algorithms. Text classification is pivotal today due to a wide range of digital documents. Multiple feature classes have been proposed for classification by numerous researchers. Genre classification tasks form the basis for advanced techniques such as native language identification, readability assessment, author identification etc. These tasks are based on the linguistic composition and complexity of the text. Rather than extracting hundreds of variables, a simple premise of text classification using only the text feature of parts-of-speech (PoS) is presented here. A new dataset gathered from Project Gutenberg is highlighted in this study. PoS analysis of each text in the created dataset was carried out. Further grouping of these texts into fictional and non-fictional texts was carried out to measure their classification accuracy using the artificial neural networks (ANN) classifier. The results indicate an overall classification accuracy of 98 and 35 % for the genre and sub-genre classification, respectively. The results of the present study highlight the importance of PoS not only as an important feature for text processing but also as a sole text feature classifier for text classification.

Keywords: speech; NLP; text; ann; genre; text types

1 Introduction

Natural language processing (NLP) is the set of methods for making the interaction between human language and computers (Eisenstein 2019). NLP is used in

***Corresponding author: Akshay Mendhakar**, Faculty of Applied Linguistics, University of Warsaw, 00-312 Warszawa, Poland, E-mail: a.mendhakar@uw.edu.pl
Darshan H S, Nitte Institute of Speech and Hearing, Mangalore, Karnataka, India

Open Access. ©2024 the author(s), published by De Gruyter on behalf of Shanghai International Studies University.
This work is licensed under the Creative Commons Attribution 4.0 International License.

almost all sectors of our lives. Text mining is an integrated method that applies NLP, machine learning and pattern classification concepts to extract meaningful information from unstructured text (Zong, Xia, and Zhang 2021). It uses linguistic concepts such as part-of-speech analysis, grammatical structures analysis, etc. (Kao and Poteet 2007). Similar to NLP, the application of text mining is evident in multiple sectors, and a few of the classical use cases have been summarized in the study by (Srivastava and Sahami 2009). Text classification is an exciting challenge to multiple researchers across various fields. With the growing body of text datasets and the development of complex methodologies to analyze them, a plethora of research papers can be noted dedicated to text classification. In order to understand the current text classification experiment, it is essential to understand what a text is and how it is classified and processed.

Generally, a text passage is a string of sentences arranged together to convey some information (Eggins 2004). It can be composed using a string of words ranging from a few to hundreds to thousands. Texts are usually connected meaningfully to narrate an event, instruct someone, or share something about it. In a literary sense, the text is any item that can be "read," regardless of whether it is a work of writing, a road sign, graffiti, or a style of apparel. In other words, it can be said that the "text" is a simple symbolic arrangement of letters/symbols (Tsapatsoulis and Djouvas 2019). From the lens of the text act, it can be defined that text forms as a means to share ideas between sender and receiver (Taruskin 1995).

Texts are written for multiple purposes, using numerous guidelines and writing structures. These written compositions are referred to as text types (Sager 1997). The terms text types and text genres are used synonymously across the literature. There are multiple methods of grouping a piece of text. Fictional and non-fictional texts are the two principle text types that can be noted across the literature (House 1997). Under these standard umbrellas of fictional or non-fictional, one can note multiple narrowly defined text types. Non-fictional texts include persuasive texts, discussions, debates, and so on.

In contrast, fictional literature includes poetry, narratives, adventures, and many more. Text types are general semantic-useful ideas (Fairclough 1992) and are not mistaken for text forms such as commercials, articles, messages, shopping records, sonnets, phone directories, books, etc. (Stubbs 1996). Text characterizations depend on the text's compositional language, the use, and distribution of punctuation, the use of spelling, and the category of a class of genre it was intended for Biber (1989). One must know the type of text it was designed to be and the characteristics of the text to outline and understand the text's ideas. Therefore, the text is composed based on the author's intent, the reader's beliefs, and rigid linguistic rules (Tsapatsoulis and Djouvas 2019).

Text types and classification can be done based on a combination of elements such as the focus on text (patterns and relationship), production (author and audience), and reception (context) (Sager 1997). Prior to discussing the linguistic compositions in fictional and non-fictional texts, it is essential to understand the core idea of "genre." Genus, which means "kind," "sort," "class," or "species," is a Latin word that has been used to describe these terms in several contexts. Cairns (1975) points out that genre classifications are as old as organized societies. As early as Greco-Roman antiquity, the classification of literary works into different genres has been a significant concern of literary theory. Genre classification has since then produced several divergent and sometimes even contradictory categories. These classifications are based on content and are empirical, not logical. They are historical assumptions constructed by authors, audiences, and critics to serve communicative and aesthetic purposes (Biber 1995). Such groupings are always in terms of distinctions and interrelations, forming a system or community of genres. The purposes they serve are social and aesthetic. Groupings arise at particular historical moments, and as they include more and more members, they are subject to repeated redefinitions or abandonment (Cohen 1986). The most prevalent classification for different types of literature consists of the triad "epic," "poetry," and "drama", according to modern literary criticism. Yet, during the seventeenth and eighteenth centuries, the conventional epic genre gave way to the novel, a new form written in prose. As a consequence, contemporary classifications now prefer using "fiction," "poetry," and "drama" as the main labels for the three major literary genres (Klarer 2013). Beyond this canonical classification, further sub-grouping is purposive and based on how an individual researcher feels about defining the same. Lee (2002) argues that even though the classification of genres is a complex and messy concept, it is essential to note the text category in terms of sub- or super-genres or just plain basic-level genres of different text types. In corpora such as British National Corpus (BNC) (Burnard 2007), International Corpus of English Great Britain (ICE-GB) corpus (Davies 2009), and Lancaster-Oslo/Bergen (LOB) corpus (Johansson, Leech, and Goodluck 1978), it can be noted that even though they follow the same canonical classification systems for basic level genre classification, minor changes in the sub- and super classification are evident. By stating the above, the focus is to highlight that deciding what a coherent genre or subgenre is can be far from easy in practice, as (sub-) genres can be endlessly multiplied or subdivided quite easily. Moreover, corpus compilers' classificatory decisions may differ from that of researchers.

However, the definition of these labels of sub-genres and supergenres is not static. In introduction to genre theory (Daniel 1997), it is highlighted that "there are no rigid rules of inclusion or exclusion. Genres are not discrete systems consisting of a fixed number of listable items." Genre definitions can differ based on society,

country, and from person to person (Daniel 1997). Even though multiple researchers follow different genre classification nomenclature. In this study, Crown's text classification system was adapted (Crown 2013). Based on Crown's classification, the text is classified into fiction, non-fictional, and poetry (these are equivalent to the primary text genres reported earlier in the paper). Each of these genres is further subdivided into mystery, adventure, explanation, or a specific form of poetry. Table 1 highlights the subgenres of the above classification. Linguistic composition varies with different text types or genres, and the combinations of the features listed solely depend on the narrative genre under usage. As the scope of our review is to analyze the linguistic composition, let us briefly review the linguistics composition reported in literature across various text types.

Table 1: Classification of text types adapted from (Crown 2013).

Fiction	Non-fiction	Poetry
Adventure	Discussion texts	Free verse
Mystery	Explanatory texts	Visual poems
Science fiction	Instructional texts	Structured poems
Fantasy	Persuasion texts	
Historical fiction	Non-chronological reports	
Contemporary fiction	Recounts	
Dilemma stories		
Dialogue, play scripts, film scripts		
Myths & legends		
Fairy tales		
Fables		
Traditional tales guidance		

2 Fictional Texts

Fictional texts are integral to textual communication and sharing stories across time. They are used to share ideas and experiences and provoke thoughts in the readers' minds. Fictional texts have a creative element intertwined with them and use literary devices for effect (Kamberelis 1999; Lee 2002). These texts focus on two most important features, the text's story, and the narrative style. These texts utilize intuitive blends of words, pictures, and sounds (Saad 2022). The fundamental part of fiction is communicating the story across time (e.g., Gardner 2000; Nussbaum 1985). Fictional texts are a source of sharing ideas and experiences while installing new emotions, thoughts, and experiences in readers' minds. In fictional texts, it can be noted that the start is usually with an opening that introduces and

establishes the set characters. A twist or hardship at the end of the text can be either a resolution/ending to the storyline (Stierle 2014). Renowned writers of fiction are unrestricted by predictable story narration structure. Many creators and narrators frequently alter or adjust the generic construction of the story by occasionally changing sequences such as time shifts, flashbacks, and backtracking. In addition, texts are sometimes enhanced with pictures (illustrations) or multimedia (pictures/video/sound). The sentence structure is written in the first or third individual (I, we, she, it, they), and utilization of past tense is common (Crown 2013). In texts of sequential sorts (plot or substance have a sequence of occasions that occurred in a specific order), fictional compositions fundamentally focus on plot and characters. To be more precise, the characteristics of the principal members, their characters, and the set structure is a vital pieces of the portrayal in every sort.

2.1 Adventure

This genre is used mainly for narrative purposes and is commonly used to retell a sequential story. The narration structure of these texts evokes excitement toward the event and results in an impactful ending (Britton and Pellegrini 2014). One of the prevalent forms of adventure texts is chronological narration. The story's main characters overcome hardships with an immense buildup of excitement and relatively little to no flashbacks. The reader adds to the story narration by building up expectations and predicting the further events of the story. The setting of the story can be any place that adds to the sense of danger and amplifies the story's impact on the reader (Wolfe 2005). Adventure texts regularly utilize various designs, permitting the readers to choose various courses through the order of events and, at times, with various goals that rely upon the decisions made by the reader. The linguistic composition of texts can be noted as a potent mix of activity, discourse, and portrayal that develops archetypical characters the reader will often think about, simultaneously moving through the plot along at a thrilling speed. The story's depiction adds to the reading experience by increasing the sense of danger or dropping clues about future events (Crown 2013). Dialogue is a component of the portrayal of the story. However, dialogues are utilized more to propel the activity than to investigate a character's sentiments or inspiration. The language used can be graded more as an artistic quality, with a robust vocabulary.

2.2 Mystery

This genre's primary purpose is to evoke the feeling of intrigue and to entertain. The text structure is frequently ordered, even in longer texts. The use of rhetorical elements often enhances the complex structural technique of the narration. However, it is prevalent with explanations and filler pieces to add more information

to the storyline (e.g., flashbacks, repeating essential information to highlight the bigger picture). Realizing and discovering what will occur can add to the anticipation (Spiegel et al. 2018). The vocabulary used adds to the mystery of the story. The most commonly used vocabulary includes odd, strange, weird, etc. The wordplay of familiar words with a hint of mysterious words, such as dark forests, uninhabited places, lonely lakes, and so forth, triggers the better experience or feeling of mystery in the readers. The use of wh-questions and modifiers, such as adjectives and adverbials, is also reasonably common. This adds to or exaggerates the mystery of the storyline (Crown 2013). Using pronouns to enhance the mystery, such as avoiding naming the characters and using common non-human pronouns such as "it," is a common practice employed by many writers of mystery texts.

2.3 Science Fiction

This genre is utilized to discuss what will come in the future (Johnson 2011). The setting is frequently a period later than the present time and may utilize structures that play with the time succession. The concepts of time travel and flashbacks are common in the sci-fi genre. Sci-fi commonly incorporates insight regarding how individuals may live, later on, foreseeing in an inventive and innovative way how innovation may progress. The linguistic structure and vocabulary usage are generally focused on adding a sense of wonderment about the storyline, and the plot usually incorporates the experience of a fast-moving lifestyle. Where modern characters are made, discourse may utilize unique structures and jargon or even elective dialects. The portrayal is critical to passing on envisioned settings, innovation, cycles, and characters (Crown 2013).

2.4 Fantasy

Fantasy texts engage us, power the creative mindset, and fuel a sense of unexplainable emotions in the reader's mind (Armitt 2005). The linguistic structure follows a mixture of simple chronological narration elements and some descriptions to enhance the fantasy. The usage of adjectives is a unique feature noted in fantasy texts and plays a vital role in the storyline. These elements describe the events and places, which, when combined with rhetorical figures such as similes and metaphors, assist the reader with envisioning what the character is experiencing. The reader travels through time in a way envisioned by the reader. The use of imagery is the most predominantly noted rhetorical figure with a description of the setting at the expense of the plot so that the actual order of things the reader has never seen. The writer tries not to make everything so fabulous that it is, on occasion, less significant or even difficult to follow but is more interested in explaining the fantastic story and setting it occurs (Crown 2013).

2.5 Myths

Mythical texts try to explain a natural event using fictional reasoning. Myths are often much longer texts than other traditional stories, such as fables. These texts are often associated with different cultures and legends that explain mysterious events and want to pass them on to further generations (Pavel 1986). Texts under this genre focus on traditions, beliefs, religion, and culture. The plot is frequently founded on a long, risky excursion, a mission, or a quest. The plot generally incorporates mind-boggling or beautiful events, where characters act superhumanly using surprising forces or superhuman creatures' assistance. They furnish a precious difference with more limited customary accounts like tales. The vocabulary used highlights the glamour and power of the character and the settings. A vivid description of characters and settings (Crown 2013). Using rhetorical figures such as imagery and symbolism to assist the readers in envisioning the characters might be standard. Similes are used to bestow a sense of wonderment and awe towards the settings. The quick portrayal of the activity keeps the plot moving along in a fast-paced manner. These texts regularly give genuine instances of using images of real-life objects and associating a symbol with it, for example, a rainbow as an image of the connection between the reality of people and the magical universe of the divine beings.

2.6 Fairy Tales

These texts were initially proposed for kids. They were stories shared to delight and pass a social message that influences good behavior. These stories are found in many societies, and many are derived from the old existing stories (Crago 2003). The plot generally starts with a vague opening, such as "Once upon a time" or "a long time ago." Language regularly mirrors the settings, previously utilizing provincial jargon and syntax. The narration follows a chronological order, and retelling events in order of occurrence is usually noted. Most stories focus on characters finding love, wealth, a home, or wisdom. In texts belonging to fairy tales, the use of supernatural power or magic is common. The end of the story follows either the theme of happily ever after or everything ending with a sad and dark twist (Tatar 2017).

2.7 Fables

A story is set to show the audience something new they should discover about existence. The story drives toward the end moral articulation, the fable's theme. For example, "hard work pays off" or "early to work reaps the reward" (Dorfman and Brewer 1994). The clear moral message at the end of the story differentiates fables from other forms of texts. The overall structure of the text is quite simple and short. The number of characters kept to a minimum follows the classical structure of

beginning, complication, and resolution. The short and basic construction of the story rules out extra subtleties of portrayal or character advancement. Discourse is used to propel the plot and focus the reader's attention on the characters. The portrayal is restricted yet explicit. Connectives are significant language features that show cause and effect, offering stability to a short narrative (Fausto 2014).

3 Non-Fictional Texts

Non-fiction texts are of different types and occur daily, even though the distinction between many different fictional text types is often blurry regarding their linguistic features and plot.

3.1 Discussion Texts

Discussion texts are focused initating or sharing a conversation or a debate on a specific topic. It provides a platform to talk about a topic in order to reach a decision or to exchange ideas (Crown 2013). Discussion texts are contemplated and adjusted outlines of an issue or disputable point. Typically, these texts plan to give at least two unique perspectives on an issue, each with elaborations, proof, and specific illustrations (Brewer and Ohtsuka 1988). The most widely recognized construction of discussion texts incorporates an assertion of the issues in question and a sense of the fundamental contentions with supporting proof/models (Bruner 1986). Another standard construction presents the contentions' for' and 'against.' Conversation texts usually end with a synopsis and an assertion of suggestion or end. The synopsis may foster one specific perspective utilizing contemplated decisions dependent on the proof given (Nystrand, Himley, and Doyle 1986). The linguistic composition of the text is made up of simple present tense and noun phrases. The use of connectives is higher due to its conversational nature. Most of the statements used in discussion texts are followed by specific illustrations using examples. The examples are occasionally augmented with multimedia such as diagrams, graphs, and other images. These additions are geared towards explaining and adding to the topic of discussion.

3.2 Explanatory Texts

Explanatory texts commonly go past a simple description of a topic; instead, they include a structured description of causes and reason or specific motives of discussions. Clarifications and reports are sometimes confused with explanatory texts (Britton and Black 2017). It might range from a simple dictionary explanation to

a detailed explanation of an event. Like all other textual content, explanatory writings are mixed with multiple other text forms to augment communication. Explanatory texts are generally used to clarify how or why or to explain something in a specific way. They build around clarifying or explaining a specific topic (Schiefele 1999). The means or stages in an interaction are clarified coherently, with appropriate examples. The linguistic complexity of the sentences is designed simply with fewer rhetorical figures. The usage of connectives is usually higher presented in the first person. It can also be noted that the sentences are simple, which does not confuse the reader. The text is interested in explaining its central theme without distracting the reader with fancy rhetorical modifications (Crown 2013). This is done using many causal connectives, such as, therefore, thus, and so on.

3.3 Instructional/Procedural Texts

As the name suggests, instructional text types are a variant of guidelines and procedures combined with specific illustrations to assist in some tasks. Text may have many visual illustrations in steps or a simple guide to keep the reader on the same stage of the narrative guide. These texts are found in all aspects of our lives and incorporate principles for games, plans, directions for making something, and bearings (Diehl and Mills 2002). These texts guarantee that something is done viably and accurately with an effective result for the user/reader. The text is quite simple and generally starts with characterizing the objective or wanted result, e.g., step-by-step instructions to make cookies, building a cabinet, etc. Charts or diagrams are regularly basic and may even replace some content. The linguistic composition includes the usage of essential action words and simple basic, easy-to-follow sentences. Guidelines may incorporate negative directions to caution the reader on specific procedure steps (Delpech and Saint Dizier 2008).

3.4 Persuasive Texts

Persuasive texts aim to share and convince a reader or audience. These texts differ based on the setting and crowd, so the persuasion is sometimes single or varied. Components of powerful composing are found in various writings, including moving picture messages and computerized sight and sound writings. A few models may incorporate proof of bias and opinions, which are presented as facts. Persuasive texts are usually composed to argue a case from a specific perspective and to motivate the audience toward a similar method of seeing things (Carrell and Connor 1991). An initial assertion (thesis) that summarizes the perspective introduced is widespread in these text types. The data presented in the texts are deliberately

coordinated to present and support a specific ideal perspective. An end articulation rehashes and supports the first proposition. The linguistic composition of the text is kept simple enough so that the reader is not lost in words but is made sure that they focus on the topic specifically. The use of simple present tense and logical connectives can be noted in these texts (Crown 2013). The use of rhetorical questions is higher, with less focus on other rhetorical figures. The usage of multimedia content is less compared to other non-fictional forms. Still, it is common to use multimedia that supports the central ideology or argument topic (To, Thomas, and Thomas 2020).

4 Poetry

Poetry represents a form of literary expression that communicates ideas, depicts scenes, or narrates tales through a condensed, lyrical presentation of language. Poems exhibit varying structures; they might include rhyming lines and meter, where the rhythm is determined by syllabic beats. Alternatively, poems can be unstructured, known as free-form, lacking any defined formal arrangement. Comparable to the tradition of spoken narratives, poetry maintains strong social and historical connections within cultures and communities. The objectives of poems are multifaceted, ranging from providing amusement, entertainment, and introspection to disseminating information, narrating stories, sharing wisdom, and preserving cultural heritage. It is important to know when rhyming is not used in poem, the unique combination of meter, imagery, and word choice, sets poem to be different from prose. The linguistic nuances present in poems can vary across different time periods and cultures, as they mirror the evolving linguistic patterns used by individuals (Crown 2013).

4.1 Free Verse

Free verse poetry is defined by its lack of a uniform rhyme scheme, metrical arrangement, or melodic structure. While these poems do possess structure, they offer considerable flexibility to poets, particularly in contrast to more strictly defined metrical forms. Poets can use varied ways of using words, use of specific style (example: informal spoken language form), short sentences and directed sentence form to maintain the attention of the readers. Free verse writings can be classified as monologue and conversation poems, where the former is written as first person single voice and the latter involves the composition of two or more voice forms.

4.2 Visual Poems

A visual poem is a type of poem intentionally crafted or arranged to be consciously perceived through visual means. Visual poems are primarily built (sometimes entirely) upon visual and auditory elements. The arrangement of words aims to construct distinct shapes, images, or communicate visual messages. The design might involve accentuating the shapes of letters, which is specifically a form of visual poems known as caligrams. Visual poems were reinvented in recent times as modern visual poems which typically utilize deconstructed elements of language. These linguistic components encompass words, petroglyphs, phonetic characters, ciphers, symbols, pictographs, iconographs, clusters, strokes, ideograms, densities, patterns, diagrams, logograms, accents, colors, and more.

5 Classification of Texts

Genre-based classification is the widely used mode of classification of text materials (Burrows 1992; Polyzou 2008; Stamatatos, Fakotakis, and Kokkinakis 2000). By understanding the genre of the literature, one can deduce the purpose and meaning of the written text. Genre helps us study the aesthetic and cultural function of language. A genre in modern life guides hundreds of readers by dividing books into different shelf spaces, making book selection more straightforward and practical. The notion of genres is straightforward in principle, yet despite several classifications (Burnard 2007; Crown 2013; Daniel 1997; Davies 2009; Johansson, Leech, and Goodluck 1978; Lee 2002) being offered, they are only partially recognized. Although the primary members of genres are typically simple to recognize, there is no inference that they are groups with well-defined limits. Our capacity to recognize and compare distinct genres aids our understanding of them.

Genre identification problem (Biber 1995; Douglas 1992; Karlgren and Cutting 1994; Kessler, Nunberg, and Schutze 1997) reveals that there are majorly three types of linguistic features, i.e., high-level (lexical and syntactic information), low-level features (token counts, character-level features), and derived features (related to word and sentence length). Word frequency statistics are extracted to measure the lexical features. These computations measure features like the frequency of stop/content words or specific counts of each pronoun (Sichel 1975; Zipf 1945, 2013). Syntactic features are computed by tagging the parts of speech involved in the text and/or grammatical features like tenses, sentence types, and so on. Features like word/sentence length, frequency of word/sentences, and so forth., are examples of character-level features, whereas ratio-related extracts from the derived features. Table 2 summarizes the most used derived linguistic features for text classification. In general, semantic and syntactic features are the most commonly

Table 2: Commonly used linguistic features in text classification experiments.

Type	Features
Low level	Average and standard deviation of sentence length
	Average and standard deviation of word length
	Average and standard deviation of token
High level	Parts of speech ratio
	Parts of speech-related measures
	Tense related features
	Sentence related features
Derived features	Additional features based on the computational scope

computed features for classification (Cao and Fang 2009; Rittman and Wacholder 2008).

The series of studies outlined further investigated effective strategies for accurate categorization. Karlgren and Cutting 1994 used a small set of textual features and discriminant analysis to predict genres – most of these features used for genre prediction involved either part of speech frequencies or text statistics. Counts based on the fixed length of texts used in their experiments were adjusted to represent frequencies rather than absolute counts. In contrast, a study Kessler, Nunberg, and Schütze (1997) underscored the emergence of genre significance within expansive and heterogeneous search spaces, although their experiment utilized insubstantial data, prompting the need for robust testing on representative datasets. Stamatatos, Fakotakis, and Kokkinakis (2000) used discriminant analysis on the frequencies of commonly occurring words. They also improved results by including frequencies of eight punctuation marks. Four identified genres from the Wall Street Journal formed the corpus. They reported 27 % errors in distinguishing four genres in 500 samples. Illouz et al. (2000) evinced the successful use of coarse level part-of-speech features in distinguishing section types from Le Monde. Their work also showed that fine-grain PoS distinctions did not make for good features in genre classification. Moreover, study Li and Liu (2003) introduced a unique text categorization algorithm utilizing positive and unlabeled data, presenting two classifications: positive and unmarked. Both positive and unlabeled data would be found in the unlabeled data. The objective was to locate labeled and unlabeled data from the unlabeled class. Their findings indicated that just one class of labeled data was utilized, implying that another class was not required to be tagged. Expanding the landscape, the study Aggarwal, Gates, and Yu (2004)'s extensive survey illuminated diverse text classification methods, offering a valuable compass for navigating various algorithms. Simultaneously, study Tong and Koller (2001) demonstrated the potency of a support vector machine and underscored the pivotal role of feature

selection. Intriguingly, study Liu et al. (2004) proposed an alternative strategy by labeling words instead of documents for classification, redefining the conventional framework. Lastly, Manning, Raghavan, and Schutze (2009) used the relative frequency of a word/term/token in texts belonging to input classes to estimate the conditional probability of that word/term/token given a class using Bernoulli Naive Bayes algorithm. Bernoulli Naive Bayes is a member of the Naive Bayes family of algorithms. It operates based on the principles of the Bernoulli Distribution and is tailored for scenarios where the features exhibit a binary nature, having values restricted to 0 or 1. In conditions, where the dataset's features follow a binary pattern, opting for the Bernoulli Naive Bayes algorithm is a fitting decision. Results revealed that Bernoulli Naive Bayes is inefficient for classifying long texts since it ignores numerous occurrences of terms. These studies underscore the significance of meticulous feature choice, algorithm suitability, dataset representation, and the nuanced interplay between approaches to accomplish precise text classification.

A study by Miltsakaki and Troutt (2008) tried to build a real-time web-based text classification system (Read X and Toreador) that can automatically analyze text difficulty. The proposed software was built using the vocabulary and word frequencies per thematic area and graded the text's readability. Text classification was based on word frequencies and was classified into the following categories: arts, career and business, literature, philosophy, science, social studies, sports, health, and technology. In the text classification experiment using Read-X, the authors reported the results of three classifiers: a MIRA (a measure of algorithm confidence), a Naïve Bayes classifier, and a maximum entropy classifier (Crammer et al. 2008). Although Read-X was created to find, categorize, and assess the reading difficulties of web material in real-time. The classifier findings indicated that the MaxEnt classifier performed at 93 % and the Naive Bayes classifier performed at 88 % for text classification. They speculated that further research could try out subclassification. They also suggested using other features in tandem with word frequencies. Further, Feldman et al. (2009) used part-of-speech histograms and principal component analysis to construct features. They propose classifying genres using the Quadratic Discriminant Analysis and Naïve Bayes algorithms. Their results highlight that computing histograms on a sliding window of five PoS tags are a suitable classification method. Similarly, Petrenz and Webber (2011) demonstrated how the accuracy of genre classification changes when each PoS feature is considered separately from other non-POS features. They conclude that PoS features should not be lumped together in genre classification experiments but evaluated individually. This is especially true for the tags VBD (past tense verb), JJ (adjective), RB (adverb), NN (singular noun), VB (base form verb), and NNP (plural proper noun). In another study, Tang et al. (2016) used a Bayesian interface in automated text categorization. They used information gain (IG) and maximum discrimination (MD) for feature selection. Their

results revealed potential usage in data mining. A study by Qureshi et al. (2019) proposed a logistic regression classifier for classifying texts into fiction or non-fiction genre. Their study highlighted a nineteen-feature classification of the Brown Corpus (Francis et al. 1982) and British National Corpus (Burnard 2007). Their results revealed 100 and 96.31 % classification accuracy, respectively. Further, they concluded that the ratio of adjectives/pronouns and adverbs/adjectives are the most significant features in classification.

Brown's corpus is a collection of annotated texts in British English classified into 15 subgenres which can be majorly classified under informative prose (non-fiction) and imaginative prose (fiction). Out of the 15 genres in this corpus, five genres of humor, editorial, lore, religion, and letters are difficult to place under either fiction or non-fiction. Out of the 500 sample texts in the brown corpus, if we exclude five genres, the sample becomes 324, which is very little to conduct any machine learning classifier for training. Similarly, in the BNC, the sample size is again too small. Therefore, there is also a need to build a better dataset for text classification problems. Recently, due to the advancement in technology, many researchers tend to utilize numerous features for classification leading to overfitting the machine learning model and lowering the performance of their models. Arguably, even though there have been reports of high-accuracy classifiers built for text classification, parts of speech estimation form the basis for numerous features in text classification.

In the present study, an attempt was made to classify texts into various genres based on the parts of speech composition. Further, a text classification experiment was planned to classify fictional versus non-fictional text genres and their subcategories based on tags of parts of speech. Poetry as a genre was excluded, as it has different compositional structure. Analysis of poetry just based on linguistic structure, does not always give a complete intended meaning of what the writer has thought.

Even though there are multiple advanced systems (Aggarwal, Gates, and Yu 2004; Biber 1995; Cao and Fang 2009; Crammer et al. 2008; Feldman et al. 2009; Francis et al. 1982; Illouz et al. 2000; Karlgren and Cutting 1994; Kessler, Nunberg, and Schütze 1997; Li and Liu 2003; Liu et al. 2004; Manning, Raghavan, and Schutze 2009; Miltsakaki and Troutt 2008; Petrenz and Webber 2011; Qureshi et al. 2019; Rittman and Wacholder 2008; Sichel 1975; Tang et al. 2016; Tong and Koller 2001; Zipf 1945; 2013) for tagging speech and grammatical units, this study approaches the problem with a more straightforward goal of building a classifier based solely on POS. This study will present a premise of grading the parts of speech statistics in each text and using the histogram function as a classification feature coupled with a quadratic discriminant classifier. Based on the above review, the following research questions were formulated for this study:

1. Is there any difference between the composition of fictional and non-fictional texts regarding parts of speech?
2. Is it possible to classify fictional and non-fictional texts based on the frequency of parts of speech?
3. Is it possible to classify texts into subgenres of fiction and non-fictional texts based on the frequency of parts of speech?
4. Can a classifier be trained to categorize texts into various genres further?

6 Methodology

The research was conducted at the LELO – Laboratory of Experimental Eye-Tracking Linguistics located at the University of Warsaw. This investigation constituted a continuous initiative to construct text-mining tool within the field of linguistic and literary sciences. This study was affiliated with the project named "Literary Text Perception and Comprehension" under the ELIT Network. The research was structured into two distinct phases; the initial phase focused on preparing the dataset and elucidating the design of the developmental tool for part-of-speech tagging.

6.1 Phase 1 of the Study: Dataset and PoS Tagging Tool Description

The study was planned in such a way that there were no participants involved in the study. Data simulations were carried out during the study using a high-processing DELL PC with an Intel i5 processor and 16 GB RAM. All the necessary data gathering and pre-processing were conducted on this PC. The data for the study was gathered from various sources. The dataset used in this study was similar to one of our previous experiments (Mendhakar 2022). For a detailed description of the corpus creation, refer to Mendhakar (2022). The linguistic composition of poems varies significantly across its types, as highlighted in the review, which would hinder the validity of the developed dataset. Therefore, only the fiction and non-fiction genres were included in the corpus.

Table 3 below summarizes the text genres chosen for this study and the number of texts selected under each heading. The selected texts were divided into chapters, and it was made sure that the overall size of each of the texts would be around 100–2000 words. The typical protocol to remove unnecessary spaces and punctuation was carried out. Therefore, at the end of phase one of the study, a dataset included pre-processed texts to remove licensing information and other text metrics that would hamper the PoS tagging of the texts.

Table 3: Summary of the dataset of the study.

Non-fiction (1514)	Fiction (2153)
1. Discussion texts (395)	5. Fairy tale (190)
2. Explanatory texts (242)	6. Fable (394)
3. Instructional texts (495)	7. Fantasy (249)
4. Persuasive texts (382)	8. Myths & legends (92)
	9. Romance (580)
	10. Science fiction (384)
	11. Mystery thriller (264)

The tool used in this study was built as part of the ongoing study on using text mining tools in literary text perception and comprehension. The final version of the tool is intended to extract numerous other features that describe the text; a part of speech tagging formed the first line of measures of the tool. The proposed tool was built using python programming language and libraries such as Spacy (Honnibal and Montani 2017) and Flask server (Grinberg 2018) for implementation.

The standard structure of the tool was to build a generalized interface where people could load their text and interact with the tool. A simple HTML landing page was designed that interacts with the app.py and communicates with the project to extract the necessary information in the loaded text. Our appy.py file had a simple and understandable structure. It includes the primary Python code that the Python interpreter will need to operate the Flask web application. For the scope of this paper, only the parts of speech tagging will be described. The spacy model initially conducts a function called sentence detection before tagging the components of speech in the provided text. This would help in determining the beginning and finishing of sentences. By using *sents* property, this step was accomplished. Tokenization was the next step in the process. We identified the essential components of a given text by tokenizing it, which then produced individual, meaningful units. These units were utilized to do more research on PoS tagging. Stop words are the most prevalent words in a specific language. Terms like *"the, are, but, and they"* are examples of stop words in the English language. Most sentences must have stop words. Stop words are typically eliminated since they are irrelevant and skew word frequency analysis in more advanced textual analysis techniques.

Lemmatization, on the other hand, is reducing inflected forms of a word while guaranteeing that the reduced form is still part of the language. A lemma is a simplified version or base word. The choice of removing stop words and lemmatization was not taken in our study, as the study aimed to analyze the text in its original form. The last step in this study phase was the parts of speech tagging. Eight basic parts of speech followed since the time of description of grammer by Dionysios

Thrax (noun, pronoun, adjective, verb, adverb, preposition, conjunction & intersection) were tagged. Also, grading the usage of numbers (numeric & string) was included. So total, each text had nine different measures. Figure 1 describes the workflow involved in extracting parts of speech-related information in the presented tool. The computed results for each text were tabulated and saved for further analysis.

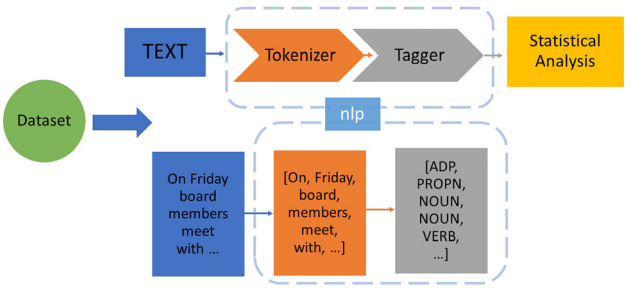


Figure 1: Workflow of parts-of-speech tagging.

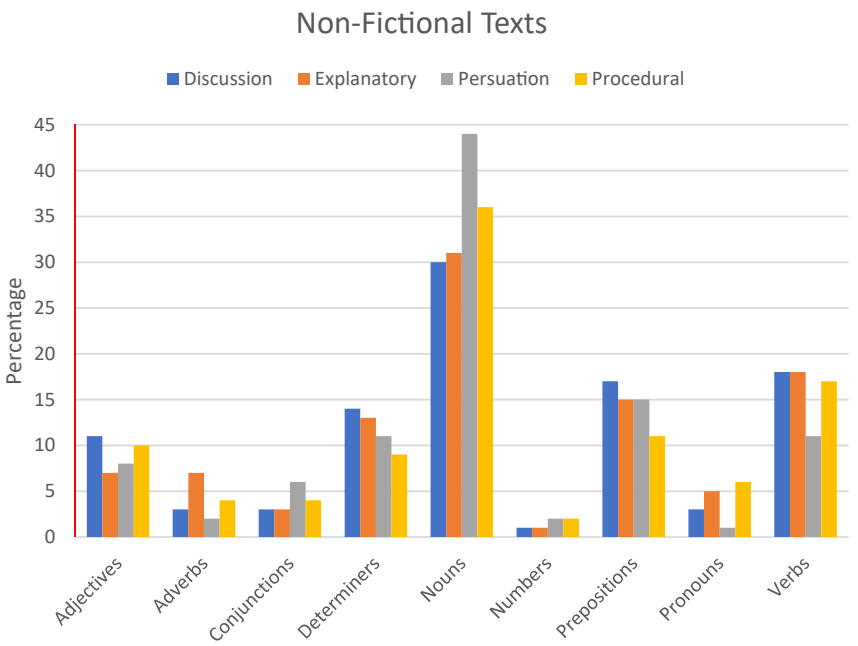


Figure 2: Percentage of PoS in each sub-genre of non-fictional text genres.

6.2 Phase 2 of the Study: Data Analysis and Testing of the Efficacy of the Classifier

The computed results of the PoS tags in each text were further subjected to statistical analysis using the SPSS software. The result section in this phase is organized to answer the first three set questions of the study. The tabulated data were subjected to a normality check using Kolmogorov-Smirnov and Shapiro-Wilk's test, and the results revealed a statistically significant Kolmogorov-Smirnov and Shapiro-Wilk's test. Therefore, the null hypothesis was accepted, concluding that the data is not normally distributed. The test results for normality did not change even after the data transformation.

The descriptive statistics of texts across fictional versus non-fictional texts were used. Overall it was noted that the mean and standard deviation of non-fictional ($N = 1514$) versus fictional texts ($N = 2153$) was 11.12 (2.774) and 11.22 (2.329), respectively. Figures 2 and 3 summarize the parts of speech noted in non-fictional and fictional texts.

The present study also wanted to classify text genres based on their PoS tags. For this task of text classification, a neural network classifier was employed. Neural networks are a popular algorithm in machine learning. They are used to solve many problems due to their versatile nature. The present study followed a feed-forward network with back-propagation following a classical network design (Gurney 2018). The designed network consisted of the PoS tags as the input with 10 hidden nodes.

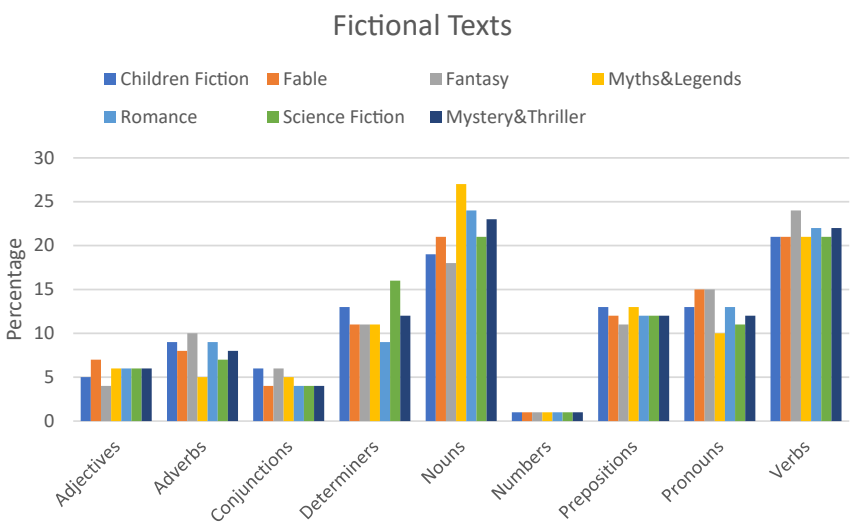


Figure 3: Percentage of PoS in each sub-genre of fictional text genres.

Multiple iterations were carried out by changing the neural network's architecture, and corresponding classification accuracies were noted.

7 Results and Discussion

The present study planned to experimentally evaluate the relative strength of PoS in texts as a standalone feature for text classification problems. The study was conducted in three phases to answer the study's research questions. Phase 1 of the study is explained in the methodological section.

7.1 Creating a New Dataset for the Text Classification Problem

Phase 1 of this study was dedicated to gathering text and building a better corpus for the text classification problem. The most popular datasets, like BNC (Burnard 2007), ICE-GB corpus (Davies 2009), and LOB corpus (Johansson, Leech, and Goodluck 1978), were compiled in the late 1970s–1990s. The latest versions of these above datasets can be noted by Burnard (2007). With a 15–20 year development time gap, the concept of genre has changed and will continue to change, as pointed out by Daniel (1997). Considering these points, developing a new dataset tailored to the experimental protocol is backed by previous research reports (Burnard 2007; Daniel 1997; Davies 2009; Johansson, Leech, and Goodluck 1978; Klarer 2013; Lee 2002). By following the canonical classification and Crown's genre classification system (Crown 2013), we provide different sub-types of fictional and non-fictional texts. In the dataset used, 2153 fictional text scripts and 1514 non-fictional scripts were used (Mendhakar 2022), roughly 40 % more representations of fictional text scripts than non-fictional ones. The 7–4 sub-type representation with fictional and non-fictional categories adds to the drawbacks of the dataset. These demographic play a crucial role in experiments like text classification using modern-day classifiers (Finch and Schneider 2006). Text scripts belonging to poetry or non-chronological reports were omitted from the dataset considered in this study. As the scope of the present study was to look at fictional and non-fictional text classification, their omission of poetry is warranted. Further, for universal adaptation of this dataset, more text scripts under each category must be included with equal representation of each text type.

Additionally, in the first phase, we described the architecture of the built PoS tagging tool. Using open source python libraries like Spacy and Flask server, the study focused on making a tool that researchers can access and modify. This effort was to support the Open Science framework (Foster and Deardorff 2017). Making the tool work as a web application that works on standard NLP libraries would

promote the accessibility of the method to researchers who are new to computational linguistic experiments.

7.2 Composition of PoS in Genres and Sub-Genres

Phase 2 of the study focused on evaluating the PoS composition across fictional and non-fictional text types and further assess the PoS of individual subgroups with respect to each other. Our review highlighted a brief description of individual text types; those claims were associated with the PoS findings and will be described in detail in the following sections.

Based on the results noted in Table 4 above, it can be observed that discernible variations in terms of standard deviations are absent between fictional and non-fictional genres. Based on Figures 2 and 3, it can be noted that there are some differences in parameters such as nouns, and verbs in non-fictional texts, etc. Similarly, other determiners and numbers in fictional texts. To explore these differences, further statistical tests were carried out.

A Mann-Whitney U test was performed to answer the first question and find differences in the composition of texts in terms of their parts of speech for fictional versus non-fictional texts. Further, this study aimed at testing the individual differences across various sub-genres of fictional and non-fictional texts; therefore, pairwise comparisons across sub-genres using independent samples Kruskal-Wallis test were performed, and Figure 4 summarizes the results obtained. Results of the Mann-Whitney U test results across the two groups noted that there is a significant difference amongst all the parameters except the adverbs ($U = 132{,}046$, $N1 = 1514$, $N2 = 2153$, $p = 0.153$, two-tailed) and verbs ($U = 2{,}489{,}826$, $N1 = 1514$, $N2 = 2153$, $p = 0.682$, two-tailed) across the two text groups of fictional and non-fictional texts. Based on the above test results, there are noticeable differences between fictional and non-fictional texts and their parts of speech. Carefully considering the percentage of different parts of speech might help us classify texts into broad genres.

7.2.1 Non-Fiction as a Group

When the non-fictional texts are evaluated for PoS elements, three groupings can be made on the concentration of PoS in these text types. Concentration of PoS denoted the amount of PoS in the given text. The first is the high-concentration PoS which are nouns (30–40 %) in non-fictional texts, mid-concentration PoS elements like verbs, prepositions, determiner and adjectives (10–20 %), and lastly, the low concentration PoS elements like pronouns, adverbs and conjunctions (<10 %). These concentrations of PoS tags. As claimed by the report of Crown (2013), support the notion that the differences in the linguistic composition are difficult to distinguish

Table 4: Mean and standard deviation scores of parts of speech for the individual genre.

Sub-genre	Adjective	Adverb	Conjunction	Determiner	Noun	Number	Preposition	Pronoun	Verb
Non-fictional	7.45 (1.59)	6.82 (2.10)	3.70 (0.95)	11.77 (2.32)	28.17 (6.52)	1.83 (2.09)	11.35 (1.58)	7.70 (3.41)	21.26 (4.41)
Discussion	8.40 (1.03)	7.00 (1.43)	4.00 (1.11)	10.80 (2.35)	27.80 (3.17)	0.80 (0.76)	12.00 (0.64)	6.80 (1.18)	22.80 (3.59)
Explanatory	7.67 (1.5)	7.66 (1.81)	3.33 (0.47)	12.66 (2.21)	27.00 (2.98)	0.83 (0.69)	12.17 (1.08)	7.17 (2.14)	21.17 (2.43)
Instructional	6.20 (1.17)	4.60 (1.37)	3.40 (0.49)	13.20 (0.75)	36.00 (2.71)	0.50 (0.64)	10.40 (1.98)	5.20 (0.75)	16.40 (1.87)
Persuasive	7.80 (1.68)	8.30 (1.43)	4.20 (1.26)	10.10 (2.04)	20.9 (4.62)	0.20 (0.4)	11.00 (1.43)	11.90 (3.88)	25.50 (3.33)
Fictional	6.62 (1.91)	7.06 (2.34)	4.65 (1.31)	12.41 (2.60)	22.49 (3.46)	0.88 (0.85)	12.82 (2.0)	12.18 (3.48)	20.92 (2.97)
Fairy tales	6.63 (1.03)	8.33 (1.06)	4.61 (0.94)	11.17 (1.57)	21.04 (1.86)	0.89 (0.36)	11.23 (1.75)	13.05 (2.04)	23.06 (2.03)
Fable	5.15 (2.54)	6.44 (2.57)	5.08 (1.73)	14.12 (3.32)	22.72 (3.49)	0.73 (1.2)	11.78 (2.76)	12.86 (3.48)	21.03 (3.56)
Fantasy	6.00 (1.53)	6.83 (1.48)	5.89 (1.33)	12.65 (2.02)	23.45 (2.86)	0.86 (0.99)	12.95 (1.61)	11.11 (3.03)	20.14 (2.81)
Mystery/thriller	6.98 (1.46)	7.10 (1.18)	3.89 (1.12)	12.17 (1.98)	21.38 (2.69)	0.92 (0.66)	13.84 (1.59)	13.23 (3.05)	20.59 (2.64)
Myths/legends	6.98 (2.28)	4.40 (1.27)	4.90 (0.94)	13.93 (1.41)	30.64 (3.62)	1.69 (0.83)	14.41 (1.36)	6.93 (2.26)	16.32 (2.1)
Romance	6.90 (1.31)	7.65 (1.40)	4.59 (0.84)	10.58 (1.97)	21.29 (2.74)	0.65 (0.65)	12.7 (1.42)	13.50 (3.04)	22.17 (2.25)
Science fiction	7.76 (1.59)	6.92 (3.78)	4.01 (0.92)	13.67 (1.68)	22.95 (2.71)	1.17 (0.64)	13.67 (1.35)	10.31 (2.97)	19.69 (2.05)

NOTE:
1. Discussion Texts
2. Explanatory Texts
3. Instructional Texts
4. Persuasive Texts
5. Children's fiction
6. Fable
7. Fantasy
8. Mystery & Thriller
9. Myths & Legends
10. Romance
11. Science Fiction

Figure 4: Comparison plots across sub-genres (Refer to Table 3 or note for codes of subgenres).

between these text types of non-fiction. As reported by To, Thomas, and Thomas (2020), the non-fictional text types are constructed to avoid distractions to the reader and keep the overall structure of the texts simple. The conversational style of discussion, explanatory and persuasive texts can be evaluated by the highest noun and connectives concentration. The parallelism of PoS tags in non-fictional text types would make classifying texts into further sub-types challenging.

7.2.2 Fiction as a Group

The concentration of PoS elements changes significantly when we look at the fictional text types. The three groupings made include different elements, i.e., the high concentration PoS elements include nouns and verbs (25 %), mid concentration PoS tags have increased and include more tags like pronouns, prepositions and determiners (10–15 %). Lastly, the low concentration PoS tags like adjectives, adverbs, and numbers (5–10 %). Additionally, when we compare the PoS tag of numbers, it can be noted that non-fictional texts use twice as much as numbers compared to fictional texts. Even though it is just 2 and 1 % respectively, this tag plays a crucial role in classification. Verbs are used to describe elements and act as a medium for explaining more about a specific event (Crown 2013). Including verbs in the high-concentration PoS tags of fictional groups supports the claims of communicative story structure of fictional texts (Gardner 2000; Nussbaum 1985). Mid-concentration PoS tags in fictional texts include pronouns, prepositions, and determiners which are used to augment and describe multiple literary devices (Kamberelis 1999).

With little differences in the PoS tags across the sub-genres of fiction, it is hard to make conclusive judgments based on PoS tags alone. However, few claims from previous studies were verified in our study. In the present study a higher concentration of conjuctions noted for fables and fairy tales emphasis the significance of story elements. This claim is inline with the reports by Fausto (2014). Science fiction is written to induce wonderment and mystery, so have more determiner PoS elements (Johnson 2011). Mythical texts have higher noun concentration as they are interested in discussing the association between specific symbols and certain beliefs (Pavel 1986). Both these claims were empirically supported in the present study. Mystery and thrillers are reported to have more adverbs, adjectives and pronouns words to induce the story's mystery (Crown 2013). These trends were not observed in the present study.

Additionally, there are a few resemblances in the PoS distributions of sub-genres like myths and legends or mystery and thrillers, fable and fairy tales, etc. These homogenous PoS findings add to the complexity of text classification and promote the subgrouping of text types. These findings align with our previous experimental findings, which showed linguistic profiling results show sub-genre overlap,

and we can group sub-genres based on complex profiling results (Mendhakar 2022). In the experiment of linguistic profiling, we used 130 parameters to come to a similar conclusion as this experiment. It is crucial to note that PoS tagging being a simple measure, can reveal similar robust findings.

7.2.3 Fiction Versus Non-Fictional Text Types

On inspection of the above results (Figure 4), the grids with green color are found to have differences, and red has no differences. Further, it can be noted that genres belonging to fictional categories, especially 7–11, are easier to classify based on specific parts of speech. In comparison, individual genres of non-fictional categories are difficult to differentiate. Only genres 7 and 9 had a perfect classification pattern based on the PoS. There are better features than prepositions for classification in further comparisons across individual non-fictional and fictional texts. Similarly, it can be noted that the subgrouping of texts in the non-fictional domain is much more complex than in the fictional category. All the calculations were made at an alpha confidence level of 0.05 % and with 10° of freedom. The overall results of the Kruskal-Wallis test suggested a statistically significant difference in terms of individual parts of speech across genres. It is possible to build a classifier by considering parts of speech as classification parameters only.

Our experiments noted that the PoS tag of numbers is one of the significant features in differentiating non-fictional (1.83 %) and functional texts (0.88 %). Similarly, pronouns were a significant feature for classification across non-fiction (7.70 %) and fiction (12.18 %). Non-fictional texts were found to be majorly noun dependent and has almost twice the amount of concentration than fictional texts (Kazmi et al. 2022). Both fictional and non-fictional texts describe a specific topic but have different nouns and verb concentrations. A comparison of nouns and verbs PoS tags suggests that non-fictional texts use more nouns to introduce a topic and explain the central element in the most precise way possible. Whereas in fictional texts, the PoS tag of verbs was almost similar to the nouns, suggesting that the story elements' description is detailed and captures the reader's attention while building immense suspense and twists (Wolfe 2005). These findings highlight the relative importance of extracted ratio measures (for example, noun/verb ratio or noun/pronoun).

In this study, an association between the overall purpose of the text and linguistic composition can be made. Even though the purpose of composing texts can vary between sharing information or evoking critical thinking, the reader is exposed to a varied concentration of PoS tags which are connected in a meaningful manner to serve the purpose of the text. Our study highlighted that these concatenations are not too different, yet they evoke a different response from the reader who reads this. This notion is supported by the speech act theory (Sbisà 2009), stating

elements of text act as a medium, but the reader and the reading context modulate the meaning.

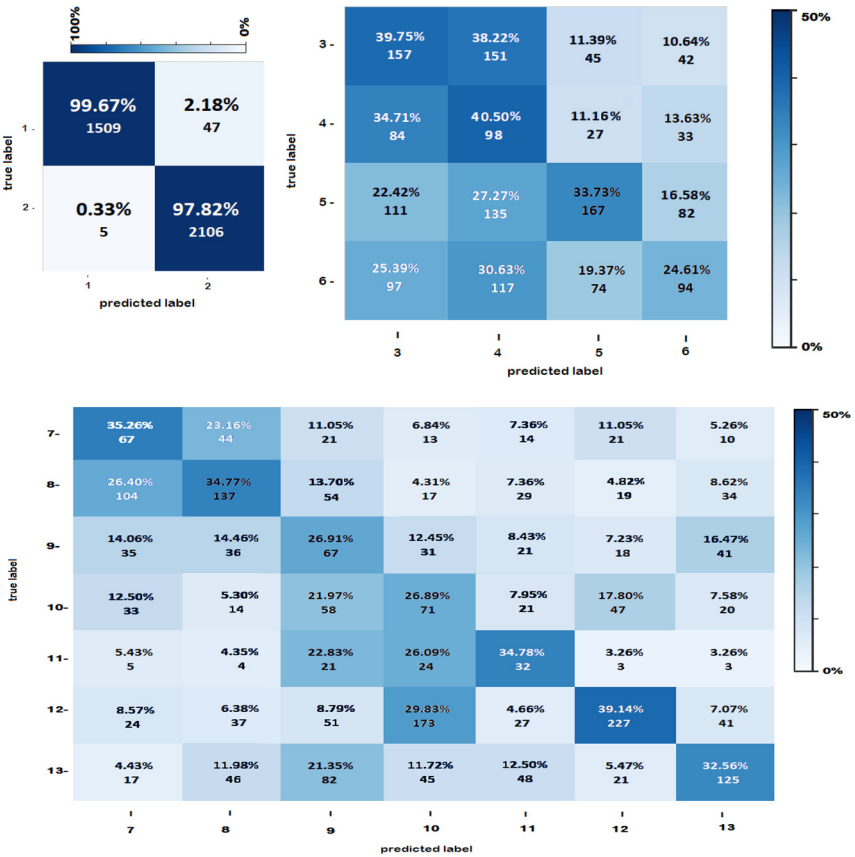
7.3 Classification Results Using ANN

When grading the overall classification accuracy of fictional versus non-fictional texts and plotting the confusion matrix, it was inferred that the neural net classification accuracy is good, with around 98 % accuracy at an F1 score of 0.98 (Figure 5). Additionally, the classification accuracy of individual fiction genres was around 33.72 %, with an F1 score ranging from 0.22 to 0.49 across the subgenres.

Multiple iterations were carried out by changing the neural network's architecture, and corresponding classification accuracies were noted. The classifier's overall performance was marginal, with no significant change in the classification accuracy. Therefore it was decided not to report these iterations.

Due to the rapid advancement of machine learning, many new features have been introduced that talk about complex classification paradigms. Various complex networks and other advanced methods have been proposed for text classification. For many researchers, genre and sub-genre-based text classification are a significant problem (Biber 1995; Burrows 1992; Douglas 1992; Karlgren and Cutting 1994; Kessler, Nunberg, and Schütze 1997; Liu et al. 2004; Manning, Raghavan, and Schutze 2009; Polyzou 2008; Stamatatos, Fakotakis, and Kokkinakis 2000; Tong and Koller 2001). It is easy to get distracted by fancy measures and complex algorithms to perform a simple task such as text classification – PoS tags from the basics of multiple advanced computational linguistic measures (Brunato et al. 2020). Hence, in this study, we used PoS tags as the feature for text classification.

We demonstrated a classification accuracy of 98 % in the genre classification task with the proposed ANN. Considering the novelty of the present study, which focused only on utilizing standard PoS tags, the classification's accuracy is exceptional compared with previous reports (Karlgren and Cutting 1994; Qureshi et al. 2019; Stamatatos, Fakotakis, and Kokkinakis 2000). The study by Stamatatos, Fakotakis, and Kokkinakis (2000) reported a 27 % error in distinguishing four genres (Wall Street Journal) in 500 samples. Karlgren and Cutting (1994) performed a series of experiments based on text classification under informative and imaginative with a classification accuracy of 96 %. A study by Qureshi et al. (2019) proposed a logistic regression classifier for classifying texts into fiction or non-fiction genre. Their study highlighted a nineteen-feature classification of the Brown Corpus (Francis et al. 1982) and British National Corpus (Burnard 2007). Their results revealed 100 and 96.31 % classification accuracy, respectively. All the above experiments are based on corpora of the 20th century but with modern classifiers. So, we note that the proposed classifier and the corpus are valid for future experiments.



NOTE: 1. Fictional texts 2. Non- fictional texts 3. Discussion texts
4. Explanatory texts 5. Instructional texts 6. Persuasive exts
7. Children fiction 8. Fable 9. Fantasy
10. Mystery& Thriller 11. Myths & Legends 12. Romance
13. Science Fiction

Figure 5: Confusion matrix plots of the overall classifier model for genre and sub-genre classification.

In the task of sub-genre classification, we reported a relatively poor classification of 35 %. Karlgren and Cutting (1994) reported a sub-genre classification using discriminant analysis of more than 20 parameters with 52 % accuracy. Similarly, the sub-genre classification of broadcast news, broadcast conversations, meetings, news, switchboard, and weblogs was reported with an accuracy of 89 % (Feldman et al. 2009). Petrenz and Webber (2011) demonstrated how the accuracy of genre classification changes by using 13 surface features (70 %) and by considering

36 PoS features (87.1 %). In their study, the dataset included news reports, educational reports, editorial reports, defense reports, and medical reports. Sub-genre classification is still challenging (Crammer et al. 2008). Compared with previous reports on sub-genre classification, the results of the present study highlight an additional spotlight on the corpus used in this study. The corpus used in the study tried to document and represent the current sub-genres of text which might add to the challenge of text classification. As the study only focused on nine core PoS tags, it is important to acknowledge that in future experiments, the 36 PoS tags can be used to repeat the experiment (Ikonomakis, Kotsiantis, and Tampakas 2005; Santorini 1990). Additionally, advanced features can be considered for the complex task of sub-genre classification.

8 Conclusion

PoS is not just a feature for text analytics but also forms the basis of estimation of multiple other NLP methodologies (Cao and Fang 2009; Rittman and Wacholder 2008). Grading the PoS across fictional and non-fictional genres helps novice researchers who are new to computational methods to connect between the established genre classification description (Crown 2013) with fundamental measure like PoS tags. Most genre classification data is often gathered via the Internet, newsgroups, message boards, and broadcast or printed news. They are multi-source, and as a result, they have a variety of formats, preferred vocabularies, and writing styles, even among texts of the same genre. The data is heterogeneous, to put it that way. Therefore research in the area of genre classification remains a challenging task.

By using a new dataset especially designed for genre classification, a detailed comparison of parts of speech across different text types was carried out in this study. Additionally, classification of subtypes of fiction and non-fiction based on PoS was carried out. Based on the results of our study, it can be noted that the fictional texts tend to have a greater adverb and adjective composition, whereas non-fictional texts are written with a higher proportion of adjective and pronoun. These distinctions aid machine learning categorization and give significant language insights. A glance at the overall mean scores across fictional and non-fictional texts reveals roughly identical scores. The usage of nouns and adjectives is roughly twice as much in non-fictional texts than in fictional texts. Similarly, adverbs, verbs, and pronouns occur twice as much in fictional texts as in non-fictional texts. These findings are backed by literature findings from many researchers (Cao and Fang 2009; Rittman 2007; Rittman and Wacholder 2008; Rittman et al. 2004).

Regarding parts of speech, grammatical classes constitute the foundation for higher emotional connotations and capture human individuality as well as their

presentation of judgments. Machine learning methods are utilized to develop reliable models for clustering, classification, and prediction. In supervised classifications, labeled text documents are utilized to categorize the text. The accuracy of these classifiers was tested on various labeled texts in this study. An ANN model with a back propagation network was used for labeled and supervised text classification. Based on the results obtained, it can be noted that the building blocks (PoS) for NLP can be used to classify texts into fictional versus non-fictional genres with high accuracies, but further subclassification of these texts into subgenres is not possible just based on PoS tags. It was also noted that differentiating fictional texts from each other is much more likely than in the non-fictional text category. The results of the present study add to the work by Burnard and McEnery (2000), Carne (1996), Cope and Kalantzis (1993), Flowerdew (1993), Hopkins and Dudley-Evans (1988), Hyland (1996), Lee (2002), and McCarthy (1998a, 1998b), to name a few, and show how a genre-based approach to analyzing texts can yield interesting linguistic insights and be pedagogically rewarding.

Even though this study focused on empirically testing whether a classifier can classify texts into genres and subgenres, further studies are required with a streamlined procedure and a more extensive dataset to arrive at a generalizable solution. The current study's major flaw is the dataset's uneven distribution while developing the corpus for the study. This issue can be addressed by considering a homogenous corpus in subsequent research. Additionally, future experiments involving different classifiers, such as Bayesian classifiers, support vector machines (SVM), k-nearest neighbor (KNN), etc., will also be planned and compared. Finding out the minimal length of a text necessary for proper categorization into fiction and non-fiction genres, as well as other relevant qualities in this respect, is an intriguing subject that we plan to pursue in the future.

Acknowledgments: I wish to thank the reviewers and Prof. Monika Płużyczka, Dr. Agnieszka Błaszczak, Dr. Niharika, Dr. Priyanka, Dr. Sonam, and Dr. Deepak for their helpful comments in improving the research study. I am also extremely grateful to all the members of IKSI at the University of Warsaw who helped me complete this research work.

Research funding: This research was conducted as part of the ELIT project, which has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie, grant agreement No 860516.

References

Aggarwal, C. C., S. C. Gates, and P. S. Yu. 2004. "On Using Partial Supervision for Text Categorization." *IEEE Transactions on Knowledge and Data Engineering* 16 (2): 245—55.

Armitt, L. 2005. *Fantasy Fiction: An Introduction*. London: A&C Black.
Biber, D. 1989. "A Typology of English Texts." *Linguistics* 27 (1): 3—44.
Biber, D. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.
Brewer, W. F., and K. Ohtsuka. 1988. "Story Structure, Characterization, Just World Organization, and Reader Affect in American and Hungarian Short Stories." *Poetics* 17 (4—5): 395—415.
Britton, B. K., and J. B. Black, eds. 2017. *Understanding Expository Text: A Theoretical and Practical Handbook for Analyzing Explanatory Text*. London and New York: Routledge.
Britton, B. K., and A. D. Pellegrini. 2014. *Narrative Thought and Narrative Language*. New York and London: Psychology Press.
Brunato, Dominique, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni 2020. "Profiling-UD: A Tool for Linguistic Profiling of Texts." In *Proceedings of The 12th Language Resources and Evaluation Conference*, 7145—51. Marseille: European Language Resources Association.
Burnard, Lou. 2007. *Reference Guide for the British National Corpus* (XML ed.). Oxford University Computing Services: Research Technologies Service. http://www.natcorp.ox.ac.uk/XMLedition/URG.
Bruner, J. S. 1986. *Actual Minds, Possible Worlds*. USA: Harvard University Press.
Burnard, L., and T. McEnery. 2000. "Genres, Keywords, Teaching: Towards a Pedagogic Account of the Language of Project Proposals." In *3rd International Conference on Teaching and Language Corpora*, 75—90. Frankfurt: Peter Lang GMBH.
Burrows, J. F. 1992. "Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information." *Literary and Linguistic Computing* 7 (2): 91—109.
Cairns, F. 1975. "Splendide Mendax: Horace Odes III. 111." *Greece & Rome* 22 (2): 129—39.
Cao, J., and A. C. Fang. 2009. "Investigating Variations in Adjective Use Across Different Text Categories." *Advances in Computational Linguistics, Journal of Research in Computing Science* 41: 207—16.
Carne, C. 1996. "Corpora, Genre Analysis and Dissertation Writing: An Evaluation of the Potential of Corpus-Based Techniques in the Study of Academic Writing." In *Proceedings of Teaching and Language Corpora* Vol. 9, (pp.127—37). Lancaster: UCREL Technical Papers.
Carrell, P. L., and U. Connor. 1991. "Reading and Writing Descriptive and Persuasive Texts." *The Modern Language Journal* 75 (3): 314—24.
Cohen, R. 1986. "History and Genre." *New Literary History* 17 (2): 203—18.
Cope, B., and M. Kalantzis. 1993. "The Power of Literacy and the Literacy of Power." In *Powers of Literacy: A Text-Type Approach to Teaching Writing*, (113, 63—89). London and New York: Routledge Taylor & Francis Group.
Crago, H. 2003. "What are Fairy Tales?" *Signal* 100: 8—26.
Crammer, K., M. Dredze, J. Blitzer, and F. Pereira. 2008. "Batch Performance for an Online Price." In *The NIPS 2007 Workshop on Efficient Machine Learning*. Vancouver, B.C., Canada: NeurIPS Proceedings.
Crown, A. 2013. *Guide to Text Types: Narrative, Non-Fiction and Poetry* [Internet]. London: National Literacy Trust. https://www.thomastallisschool.com/uploads/2/2/8/7/2287089/guide_to_text_types_final-1.pdf (accessed December 5, 2021).
Daniel, C. 1997. *An Introduction to Genre Theory* [Internet]. http://www.aber.ac.uk/media/Documents/intgenre/chandler_genre_theory.pdf (accessed December 5, 2021).

Davies, M. 2009. "The British Component of the International Corpus of English (ICE-GB), Release 2, and: Diachronic Corpus of Present-Day Spoken English (DCPSE), and: The International Corpus of English Corpus Utility Program (ICECUP), Version 3.1." *Language* 85 (2): 443—5.

Delpech, E., and P. Saint Dizier. 2008. "Investigating the Structure of Procedural Texts for Answering How-To Questions." In *Language Resources and Evaluation Conference (LREC 2008)*, 544—550. Morocco: European Language Resources Association (ELRA).

Diehl, V. A., and C. B. Mills. 2002. "Procedural Text Structure and Reader Perceptions and Performance." *The Journal of General Psychology* 129 (1): 18—35.

Dorfman, M. H., and W. F. Brewer. 1994. "Understanding the Points of Fables." *Discourse Processes* 17 (1): 105—29.

Douglas, D. 1992. "The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings." *Computers and the Humanities* 26 (5): 331—45.

Eggins, S. 2004. *Introduction to Systemic Functional Linguistics*, 2nd ed. London: Continuum International Publishing Group.

Eisenstein, J. 2019. *Introduction to Natural Language Processing*. Cambridge, MA: MIT Press.

Fairclough, N. 1992. "Discourse and Text: Linguistic and Intertextual Analysis Within Discourse Analysis." *Discourse & Society* 3 (2): 193—217.

Fausto, F. M. 2014. "Linguistic and Multimodal Perspectives on the Fable." Doctoral diss. Belfast: Queen's University.

Feldman, S., M. A. Marin, M. Ostendorf, and M. R. Gupta. 2009. "Part-of-Speech Histograms for Genre Classification of Text." In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4781—4784. Taipei, Taiwan: IEEE.

Finch, W. Holmes, and Mercedes K. Schneider. 2006. "Misclassification Rates for Four Methods of Group Classification: Impact of Predictor Distribution, Covariance Inequality, Effect Size, Sample Size, and Group Size Ratio." *Educational and Psychological Measurement* 66 (2): 240—57.

Flowerdew, J. 1993. "An Educational, or Process, Approach to the Teaching of Professional Genres." *ELT Journal* 47 (4): 305—16.

Foster, Erin D., and Ariel Deardorff. 2017. "Open Science Framework (OSF)." *Journal of the Medical Library Association: JMLA* 105 (2): 203.

Francis, W. N., H. Kucera, H. Kučera, and A. W. Mackie. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Gardner, J. 2000. *On Moral Fiction*. New York: Basic Books.

Grinberg, M. 2018. *Flask Web Development: Developing Web Applications With Python*. Sebastopol: O'Reilly Media, Inc.

Gurney, Kevin. 2018. *An Introduction to Neural Networks*. London: CRC Press.

Honnibal, M., and I. Montani. 2017. "spaCy 2: Natural Language Understanding With Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing." *To Appear* 7 (1): 411—20.

Hopkins, A., and T. Dudley-Evans. 1988. "A Genre-Based Investigation of the Discussion Sections in Articles and Dissertations." *English for Specific Purposes* 7 (2): 113—21.

House, J. 1997. *Translation Quality Assessment: A Model Revisited*. Tübingen: Gunter Narr Verlag.

Hyland, K. 1996. "Talking to the Academy: Forms of Hedging in Science Research Articles." *Written Communication* 13 (2): 251—81.

Ikonomakis, M., S. Kotsiantis, and V. Tampakas. 2005. "Text Classification Using Machine Learning Techniques." *WSEAS Transactions on Computers* 4 (8): 966—74.

Illouz, G., B. Habert, H. Folch, S. Fleury, S. Heiden, P. Lafon, and S. Prevost. 2000. "TyPex: Generic Feature for Text Profiler." In *RIAO*, 12—14. France: College de France.

Johansson, S., G. N. Leech, and H. Goodluck. 1978. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use With Digital Computer*. Norway: Department of English, University of Oslo.

Johnson, B. D. 2011. “Science Fiction Prototyping: Designing the Future With Science Fiction.” *Synthesis Lectures on Computer Science* 3 (1): 1—90.

Kamberelis, G. 1999. “Genre Development and Learning: “Children Writing Stories, Science Reports, and Poems”.” *Research in the Teaching of English* 33: 403—60.

Kao, A., and S. R. Poteet. eds. 2007. *Natural Language Processing and Text Mining*. London: Springer Science & Business Media.

Karlgren, J., and D. Cutting. 1994. “Recognizing Text Genres With Simple Metrics Using Discriminant Analysis.” In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Kazmi, Arman, Sidharth Ranjan, Arpit Sharma, and Rajakrishnan Rajkumar. 2022. “Linguistically Motivated Features for Classifying Shorter Text into Fiction and Non-Fiction Genre.” In *Proceedings of the 29th International Conference on Computational Linguistics*, 922—937. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Kessler, B., G. Nunberg, and H. Schütze. 1997. “Automatic Detection of Text Genre.” In *In 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid: Association for Computational Linguistics.

Klarer, M. 2013. *An Introduction to Literary Studies*. New York: Routledge.

Lee, D. Y. 2002. “Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle.” In *Teaching and Learning by Doing Corpus Analysis*. Leiden: Brill.

Li, X., and B. Liu. 2003. “Learning to Classify Texts Using Positive and Unlabeled Data.” In *IJCAI’03: Proceedings of the 18th international joint conference on Artificial intelligence*, 587—592. United States: Morgan Kaufmann Publishers Inc.

Liu, B., X. Li, W. S. Lee, and P. S. Yu. 2004. “Text Classification by Labeling Words.” In *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence*, 425—430. San Jose, California, USA: AAAI Press / The MIT Press 2004.

Manning, C. D., P. Raghavan, and H. Schütze. 2009. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

McCarthy, M. 1998a. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

McCarthy, M. 1998b. “Taming the Spoken Language: Genre Theory and Pedagogy.” *The Language Teacher* 22 (9).

Mendhakar, A. 2022. “Linguistic Profiling of Text Genres: An Exploration of Fictional Versus Non-Fictional Texts.” *Information* 13 (8): 357.

Miltsakaki, E., and A. Troutt. 2008. “Real Time Web Text Classification and Analysis of Reading Difficulty.” In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, 89—97. Columbus, Ohio: Association for Computational Linguistics.

Nussbaum, M. 1985. ““Finely Aware and Richly Responsible”: Moral Attention and the Moral Task of Literature.” *The Journal of Philosophy* 82 (10): 516—29.

Nystrand, M., M. Himley, and A. Doyle. 1986. “The Structure of Written Communication: Studies in Reciprocity Between Writers and Readers.” In *The Structure of Written Communication*. Orlando, Tokyo: Academic Press.

Pavel, T. G. 1986. *Fictional Worlds*. Cambridge, Massachusetts, and London: Harvard University Press.

Petrenz, P., and B. Webber. 2011. “Stable Classification of Text Genres.” *Computational Linguistics* 37 (2): 385—93.

Polyzou, A. 2008. “Genre-Based Data Selection and Classification for Critical Discourse Analysis.” In *Papers from the Lancaster University Postgraduate Conference in Linguistics and Language Teaching*, Vol. 2, 104—35. Lancaster: Lancaster University.

Qureshi, M. R., S. Ranjan, R. Rajkumar, and S. Kushal. 2019. “A Simple Approach to Classify Fictional and Non-Fictional Genres.” In *Proceedings of the Second Workshop on Storytelling*. Florence, Italy: Association for Computational Linguistics.

Rittman, R. J. 2007. *Automatic Discrimination of Genres: The Role of Adjectives and Adverbs as Suggested by Linguistics and Psychology*. New Jersey: Rutgers The State University of New Jersey-New Brunswick.

Rittman, R., and N. Wacholder. 2008. “Adjectives and Adverbs as Indicators of Affective Language for Automatic Genre Detection.” In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, Vol. 2, 65—72. University of Aberdeen: The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Rittman, R., N. Wacholder, P. Kantor, K. B. Ng, T. Strzalkowski, and Y. Sun. 2004. “Adjectives as Indicators of Subjectivity in Documents.” *Proceedings of the American Society for Information Science and Technology* 41 (1): 349—59.

Saad. 2022. *Why Classics in Literature Stand the Test of Time* [Internet]. Dailyo.in. https://www.dailyo.in/arts/classics-english-literature-thomas-hardy/story/1/22929.html (accessed December 5, 2021).

Sager, J. C. 1997. “Text Types and Translation.” *Benjamins Translation Library* 26: 25—42.

Santorini, Beatrice. 1990. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*.

Sbisà, Marina. 2009. “Speech Act Theory.” *Key Notions for Pragmatics* 1: 229—344.

Schiefele, U. 1999. “Interest and Learning from Text.” *Scientific Studies of Reading* 3 (3): 257—79.

Sichel, H. S. 1975. “On a Distribution Law for Word Frequencies.” *Journal of the American Statistical Association* 70 (351a): 542—7.

Spiegel, S., B. Beil, H. Schwaab, and D. Wentz. 2018. “The Big Genre Mystery—The Mystery Genre.” *LOST in Media* 19: 29.

Srivastava, Ashok N., and Mehran Sahami. eds. 2009. *Text Mining: Classification, Clustering, and Applications*. FL: CRC Press.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis. 2000. “Text Genre Detection Using Common Word Frequencies.” In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Stierle, K. 2014. “The Reading of Fictional Texts.” In *The Reader in the Text*, 83—105. Princeton: Princeton University Press.

Stubbs, M. 1996. *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.

Tang, B., H. He, P. M. Baggenstoss, and S. Kay. 2016. “A Bayesian Classification Approach Using Class-Specific Features for Text Categorization.” *IEEE Transactions on Knowledge and Data Engineering* 28 (6): 1602—6.

Taruskin, R. 1995. *Text and Act: Essays on Music and Performance*. USA: Oxford University Press.

Tatar, M. 2017. *The Classic Fairy Tales (Second International Student Edition) (Norton Critical Editions)*. New York: WW Norton & Company.

To, V., D. Thomas, and A. Thomas. 2020. “Writing Persuasive Texts: Using Grammatical Metaphors for Rhetorical Purposes in an Educational Context.” *Australian Journal of Linguistics* 40 (2): 139—59.

Tong, S., and D. Koller. 2001. “Support Vector Machine Active Learning With Applications to Text Classification.” *Journal of Machine Learning Research* 2: 45—66.

Tsapatsoulis, N., and C. Djouvas. 2019. “Opinion Mining from Social Media Short Texts: Does Collective Intelligence Beat Deep Learning?” *Frontiers in Robotics and AI* 5: 138.

Wolfe, M. B. 2005. “Memory for Narrative and Expository Text: Independent Influences of Semantic Associations and Text Organization.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31 (2): 359.

Zipf, G. K. 1945. “The Meaning-Frequency Relationship of Words.” *The Journal of General Psychology* 33 (2): 251—6.

Zipf, G. K. 2013. “Selected Studies of the Principle of Relative Frequency in Language.” In *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA and London, England: Harvard University Press.

Zong, C., R. Xia, and J. Zhang. 2021. *Text Data Mining*, Vol. 711, 712. Singapore: Springer.