Identifying Character Personas Using Natural Language Processing

Liana Nunziato

June 11, 2015

Abstract

Understanding the persona of a character can help readers better understand the storyline of a novel, which can be especially important when studying novels from different time periods. Generally, a reader generates a persona by looking at a character's actions, dialog, behavior, opinions, and relationships. This project produced a program that generates character personas in a novel by displaying the most frequent verbs and adjectives that are in the same sentence as a character's name. Using those words, the program also generates a positive or negative rating for each chapter, which is displayed in a graph of the entire book.

Contents

1 Introduction **4**

2 Background and Related Work **8**

3 Approach **9**

4 Results **11**

4.1 Character 1: Marianne Dashwood from *Sense and Sensibility* 11

4.2 Character 2: Elinor Dashwood from *Sense and Sensibility* 14

4.3 Character 3: John Willoughby from *Sense and Sensibility* 15

4.4 Character 4: Edward Ferras from *Sense and Sensibility* 16

4.5 Character 5: Colonel Brandon from *Sense and Sensibility* 16

4.6 Character 6: Emma Woodhouse from *Emma* 17

4.7 Character 7: Mr. Knightley from *Emma* 17

4.8 Character 8: Harriet from *Emma* 19

4.9 Comparative Analysis 20

5 Evaluation **21**

5.1 *Sense and Sensibility* Data 21

5.2 *Emma* Data 24

6 Conclusion **25**

7 Bibliography **26**

8 Appendix **27**

List of Figures

1 Graph of Polarity for Marianne throughout *Sense and Sensility*. 11

2 Graph of Polarity for Elinor throughout *Sense and Sensibility* 12

3 Graph of Polarity for Willoughby throughout *Sense and Sensibility* 12

4 Graph of Polarity for Edward throughout *Sense and Sensibility* 13

5 Graph of Polarity for Brandon throughout *Sense and Sensibility* 13

6 Graph of Polarity for Emma throughout the novel *Emma* 18

7 Graph of Polarity for Knightley throughout the novel *Emma* 18

8 Graph of Polarity for Harriet throughout the novel *Emma* 19

1 Introduction

Authors make precise decisions when developing the persona of their characters. The Merriam–Webster dictionary defines a persona as "the way you behave, talk, etc., with other people that causes them to see you as a particular kind of person: the image or personality that a person presents to other people." In other words, a persona can be identified in a character's actions, dialog, and the way that they interact with others. Writers use the descriptions of a character's actions to give us insight to their emotions, as well as their state of mind (Zunshine 4). Writers carefully develop the actions, dialog and behavior of a character to accurately reflect that character's persona so that readers can understand the events of a story line.

In identifying the various personas of the characters in a novel, readers can gain a better insight of that novel, in particular the plot line of a story. By understanding a persona, readers understand the character's actions and therefore can further analyze a story. Currently, this work is done manually as the readers process each sentence themselves. Readers interpret the emotions or characteristics of a particular character based on the actions of that character. This can be seen in many of Jane Austen's works including her most notable novel, *Pride and Prejudice*. In the novel, when Mr. Darcy and Elizabeth are eating dinner with her family, the book reads, "Darcy only smiled; and the general pause which ensued made Elizabeth tremble lest her mother should be exposing herself again." It is obvious to the reader the emotions Elizabeth has regarding the encounter. Elizabeth is not trembling because she is cold, but because she is embarrassed about her mother. The awkward moment that occurs at the dinner causes Elizabeth discomfort and results in her trembling. Currently, readers have to manually break down her emotions through reading below the surface level.

The sentence structure that Jane Austen creates for each of her characters has been shown to be very deliberate and unique. The work of Mooneyham (1998) illustrates that Jane Austen was very deliberate in her linguistic techniques for the main characters in *Pride and Prejudice*. Mooneyham states, "The first half of the novel displays the growing linguistic divisions between Elizabeth and Darcy as each perceives reality according to his or her own habit of speech-Elizabeth through wit and its attendant blindness, Darcy through the language of reserve and privilege" (46). Said another way, the style of words that the characters used when they spoke were unique and associated with who they were as a character. This proves that

Austen purposefully uses words to portray her characters in a certain way. This causes the plot line to develop dependent on her character's personas. The personas of each chapter and the way they interact with each other is a statement on the way society was during the time period the novel was set in or written in. This is important to note because the linguistic dialog that has been proven by Mooneyham to be unique to characters means that if a character's name is in a sentence before or after their dialog, the words in that sentence will be relevant to their persona.

The deliberate word choice that Jane Austen uses in *Pride and Prejudice* is seen in all of her works including *Sense and Sensibility* and *Emma*, which are the texts that I will be used in this paper to discuss the program that has been produced. *Sense and Sensibility* is a novel about two sisters, Elinor and Marianne, where the novel *Emma* is about a somewhat selfish girl named Emma Woodhouse.

Although readers are able to do most of the analysis of a persona manually, through examining the character's dialog as well as adjectives used to describe a character, some of the attributes are more difficult to pick up on. Further complications can arise when tensions exist between characters, such as the notable relationship that Elizabeth Bennet and Mr. Darcy have. In the beginning of the *Pride and Prejudice*, many characters describe Mr. Darcy as judgmental and proud, and readers hear him use negative words to describe Elizabeth. Therefore, his persona is portrayed very negatively.

The opinions of a character, the relationships that character has, the sentiment that the character uses and the sentiment that is used when describing the character is helpful in developing a persona. These elements are much more difficult for a reader to pick up on. If taken into account his actions, such as his smile, and positive thoughts and opinions toward Elizabeth, his persona is more well rounded.

This paper describes a tool that could be used to identify a persona in ways that are difficult for a reader to pick up on. This could be useful in looking at the different characters in a novel, and assist the reader in developing an understanding of a character's persona. Not only would the tool assist in helping a reader decide if a character is innately good or bad but this proposed tool might also be helpful in studying how a particular character might have changed throughout a novel. Comparing the way Mr. Darcy's persona is portrayed in the beginning of *Pride and Prejudice* will most likely differ from the way his persona is portrayed at the end of the novel when he is thought of fondly by Elizabeth Bennett and her family.

This is similar to comparing the way that Willoughby, a charismatic man, is thought of positively by other characters in the beginning of *Sense and Sensibility* to where he is thought of negatively at the end of the novel. In other words, the way both characters are viewed changes throughout the course of the novel.

This paper suggest a tool that looks at the adjectives and verbs associated with a character as well as determining the polarity, or whether a word is positive or negative for each of those verbs and adjectives.

In the first chapter of *Sense and Sensibilty*, a sentence reads,

> "Elinor, this eldest daughter, whose advice was so **effectual**, *possessed* a **strength** of *understanding*, and **coolness** of judgment, which **qualified** her, though only nineteen, to *be* the counsellor of her mother, and *enabled* her frequently to *counteract*, to the advantage of them all, that **eagerness** of mind in Mrs. Dashwood which must generally *have led* to *imprudence*."

This sentence is associated with the character name Elinor, which is highlighted in red. The blue italicized represents the verbs associated with the character name and the bold underlined words represents the adjectives associated with the character name. The Present Tense Verbs for this sentence are: be, counteract, have, imprudence. The Past Tense Verbs for this sentence are: was, enabled. The Past Participle Verbs: possessed, qualified, nineteen, led. The Adjectives for this sentence are: effectual, strength, coolness, qualified, eagerness. Present tense is referring to an action that is happening it current time. Past tense verbs are verbs that are done in the past. Past Participle verbs are also verbs that discuss past actions but in the form of irregular verbs, such as written. Adjectives are attributes. Looking at these lists separately may help in understanding how a character is developing over the course of the book. Past tense actions may be different than present tense actions if a character is developing.

After pulling the adjectives and verbs associated with the character name, I remove the stop words from the list. Stop words are words that are very common in the English language, and would not help in developing a character's persona. A few common stop words are: the, is, at, which, and on. I also make sure that the adjectives and verbs are in the subjectivity Lexicon. The work of Wilson, et al. (2005) produced this existing lexicon that has words with a part of speech tag (POS) and a polarity number associated with them. Part of speech are the different types of words that exist, such as nouns, verbs, adjectives, adverbs, pronouns, conjunctions, prepositions and interjections. In the lexicon, each word has a negative, positive,

or neutral polarity, or rating. I associated these different polarities with numbers that are either 1, -1, or 0. For instance abandoned would be a 'VBD' or past tense word and it would be a negative polarity, or -1. If a character is associated with the word abandon, that character would accrue a -1 toward their overall polarity. It is my hope that the tool will produce information about a character's persona and that the information is helpful to the readers.

If there is a drastic change in a character, this could potentially produce a words that are frequently associated with a character, but do not have to do with the character's persona at the end of the novel. This is something that needs to be investigated after the data is collected for multiple books. Another potential problem is that sentences could be associated with multiple characters. For instance, in the above example, Mrs. Dashwood's name appears. If compiling a persona for her using the suggested tool, this sentence would be included. The sentence contains words that are not relevant to her persona and could skew her overall data.

It is essential to understand these character personas to fully understand the mentality behind their actions and the story line. Novels, although fictional, give readers huge insights into the time periods in which they were written and the major ideas and influences of the general public. Through studies of English literature, many have gained a better understanding of the ideas that generate the creative writings or technological developments of a certain time period. Understanding the persona of a character can help us truly understand a text, which is imperative for understanding the past.

This paper will discuss the current work that is being done to look into character's personas and identifications, and the approach that I took in developing a tool that would generate a persona.

The program developed is studied with two selected novels whose character's personas have been thoroughly studied manually. These novels are *Sense and Sensibility* and *Emma* by Jane Austen. Therefore, the results of the program can be evaluated on accuracy of a character's persona, as well as if the results can assist in the analysis of the characters persona.

2 Background and Related Work

As previously mentioned, many English literature researchers have analyzed Austen's work to look at sentence structure and voice. In Mooneyham's work, the unique styles that exist for two different characters was analyzed and it was determined that the two different characters had different ways of talking. These unique voices may be relatable to the polarity that exists for each character. For instance, if a character is overall very positive and uses more positive words when speaking compared to another character in the same novel, it might be reflected in the polarity graph. Mooneyham's work shows that because there is a stylistic differenece and that word choice is unique to a character, that this is entirely possible.

Some recent work has been on identifying personas based on the stereotypical actions of a character as well as attributes that are used to describe a character. Bamman, O'Connor and Smith used agent verbs, patient verbs and attributes to identify personas of characters in movie plot summaries. They have presented models that learn words to topics and topics to personas. For example, the word 'strangle' is mostly likely an 'assault' word, 'assault' is generally done by a 'villain'. These identifiers show that words associated with a character can be telling of that character's persona. It is my hope that if similar words are continuously associated with a character throughout a novel, it will reflect in the adjectives and verbs associated with the character. For instance, a serial killer may assault many people in a novel, then that action is frequently associated with the characters name and is illustrated in the data produced for the user of this program.

Further work has been researched by Wilson et al. (2005) in which a new approach to phrase-level sentiment analysis is suggested. This new approach is able to determine the polarity of an expression, meaning whether the polarity of a word that has a certain part of speech tag is positive, negative or neutral. This work is extremely important to the program that has been developed because the polarity for the actions and verbs associated with a character's name may be relevant to said character's persona.

3 Approach

As discussed in the Introduction, a character's persona can be identified through adjectives, verbs (actions), dialog, and emotions. My program will focus on adjectives and verbs, as well as polarity. After selecting a novel and characters, I pull words with the individual character's name and part of speech tag those words. I then extract the verbs and adjectives from the list of words associated with a character, and display the most frequent verbs and adjectives. The verbs and adjectives are also used to develop a polarity count for a character through the novel. The rest of the section will explore each step more thoroughly:

1. Selecting a Novel

   I use novels from Project Gutenberg, which is a free online library of digitized books. See Appendix A to get directions on how to use a novel with my program.

2. Part of Speech Tagging

   By using an existing tool, I am able to look at an entire novel from Project Gutenberg and tag the part of speech (POS) of each word in that novel. This includes adjectives, verbs and nouns.

3. Descriptions (adjectives)

   By pulling sentences that characters name, I identify frequent adjectives associated with a character. It is hypothesized that the adjectives that are most frequently associated in the same sentence as a characters name will relate to a description of said character's persona.

4. Actions (verbs)

   Similarly, the actions (verbs) that occur in a sentence with a character's name may be used to identify a character's persona. It is hypothesized that verbs that are most frequently associated with a character's name will be telling of their personality.

5. Polarity

   At this point adjectives and verbs would be associated with a character name. This would be similar to the example that was used in the Introduction where the verbs and adjectives for a sentence containing the name Elinor was extracted. I use an existing corpus called Subjectivity Lexicon that has

tagged words with polarity, which I have modified (See Appendix B). With this modified Lexicon, I search with the extracted adjectives and verbs that I have obtained for a character and attach a polarity to those words. With each word associated with a character, there is now a value that can compute to the overall polarity for a chapter, or for the entire novel.

For the example that was in the Introduction, this would be the words that were associated with Elinor:

Present Tense Verbs 'VB': have, imprudence

Past Tense Verbs 'VBD': possessed, led

Adjectives: effectual, strength, coolness, qualified, eagerness

The red words represent positive words and the blue words represent negative words. There are 5 positive and 2 negative words in this sentence. Therefore, Elinor's overall polarity for this sentence would be 3. Notice that the Past Tense words that were in the example from the Introduction have been removed. This is because these were either stop words, which are words commonly used in the English language, or because there were words that were not in the Subjectivity Lexicon. The words that are shown here has either a neutral polarity or a polarity that is positive or negative depending on the context of the rest of the sentence.

It is my hope that the polarity of a character is associated with their persona. Since this program uses adjectives and verbs to develop the polarity it is my hope that the polarity will be accurate.

I give the user the following output:

- Most Frequent Present Tense Verbs, which are tagged as 'VB'
- Most Frequent Past Participle Verbs, which are tagged as 'VBN'
- Most Frequent Past Tense Verbs, which are tagged as 'VBD'
- Most Frequent Adjectives

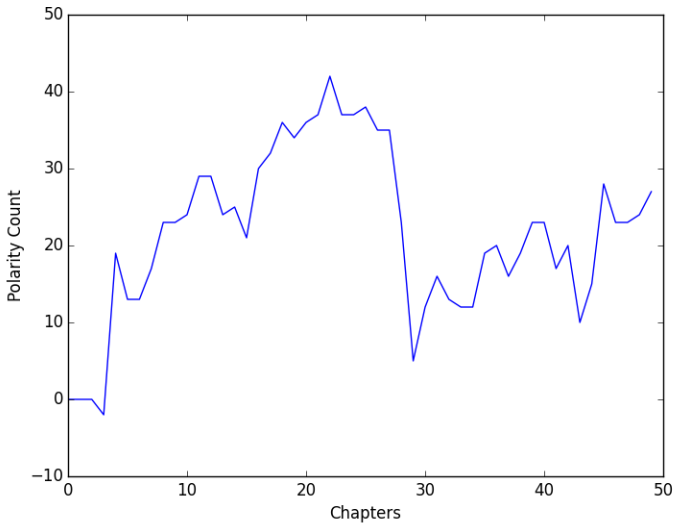

Figure 1: Graph of Polarity for Marianne throughout *Sense and Sensility*.

- The polarity change throughout the novel for the character in a graph form.

With each most frequent word, I plan to also give the user the amount of times that word was associated with a character's name.

It is hypothesized that the combination of the identifiers discussed in the approach section can be used to accurately identify a character's persona. It is thought that keeping these tagged words separate, it might help the user in understanding a character's persona and how it might be changing throughout a novel. In other words, past actions would be different than present tense actions depending on how a character is changing. It would also be interesting to note which tagged group was most relevant to a character's persona.

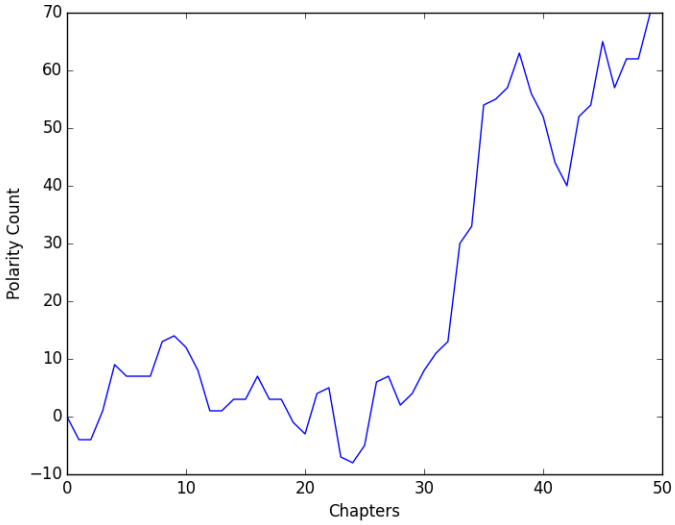

Figure 2: Graph of Polarity for Elinor throughout *Sense and Sensibility*

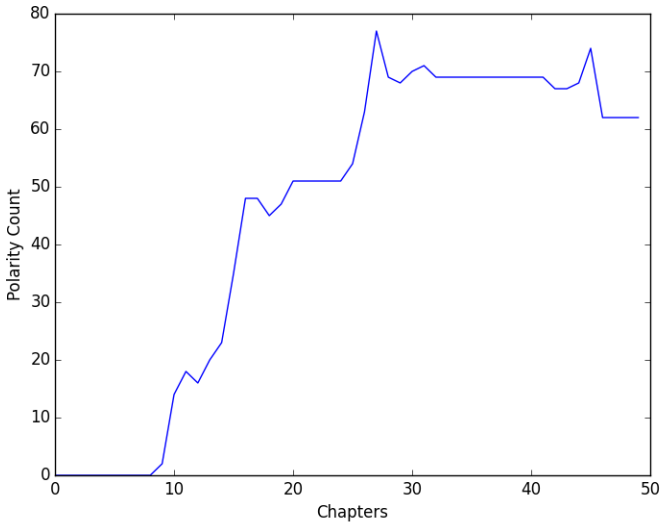

Figure 3: Graph of Polarity for Willoughby throughout *Sense and Sensibility*

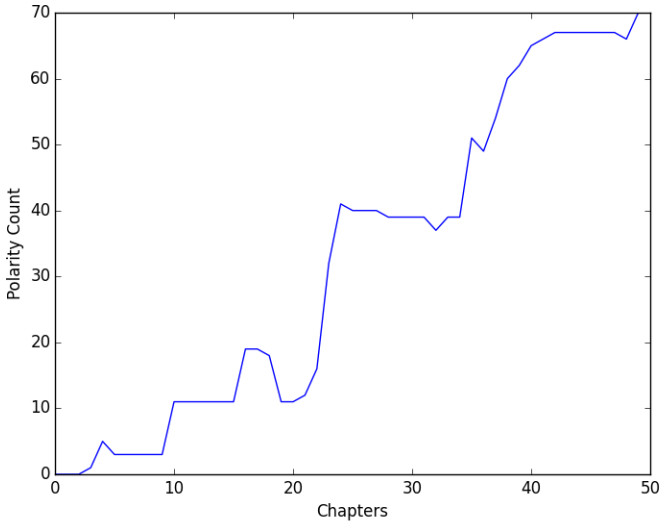

Figure 4: Graph of Polarity for Edward throughout *Sense and Sensibility*

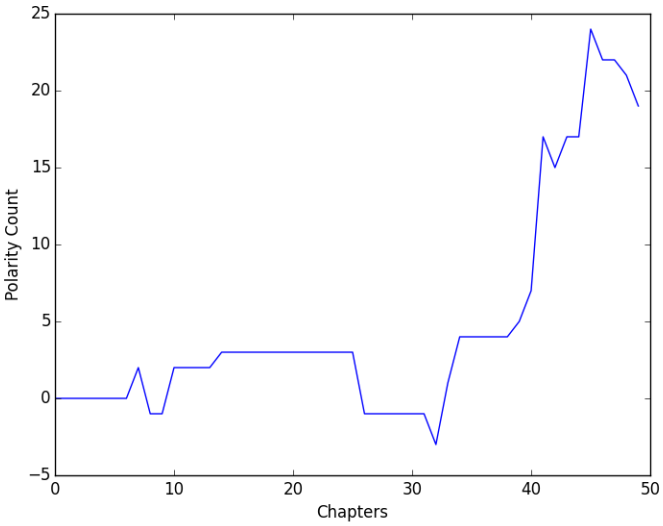

Figure 5: Graph of Polarity for Brandon throughout *Sense and Sensibility*

4 Results

4.1 Character 1: Marianne Dashwood from *Sense and Sensibility*

Marianne Dashwood is the middle daughter of Mrs. Dashwood, a widower who was the second wife to Mr. Drashwood. Marianne is seventeen in the novel. Her sister, Elinor and her are both unmarried in the beginning of the novel and are looking to marry well because their father's estate was given to his son from his first marriage.

Frequent Words for Marianne

Present tense Verbs 'VB': [('hope', 12), ('help', 9), ('excuse', 9), ('think', 8), ('gain', 8)]

Past Tense Verbs 'VBD': [('thought', 32), ('felt', 19), ('knew', 9), ('fell', 3), ('resolved', 2)]

Past Participle Verbs 'VBN': [('thought', 6), ('felt', 6), ('mistaken', 5), ('pleased', 4), ('hurt', 4)]

ADJ: [('great', 21), ('much', 20), ('good', 14), ('happy', 14), ('sure', 12)]

Words from the data produced that may be relevant to Marianne's persona: hope, help, excuse, think, gain, thought felt, knew, mistaken, pleased, hurt, happy, sure. In other words, most of the words produced could be related to Marianne's persona on some level. However, hope, felt, mistaken, pleased and hurt seem the most relevant to Marianne's character.

As seen in Figure 1, the polarity for Marianne fluctuates throughout the novel with a general trend of increasing until a significant dip prior to Chapter 30. The graph then continues to have a general increase in polarity. The negative polarity that occurs for Marianne prior to Chapter 30 is associated with her love interest, Willoughby leaving town.

4.2 Character 2: Elinor Dashwood from *Sense and Sensibility*

Elinor Dashwood is the older sister of Marianne and she is nineteen. Elinor is the main character in the novel and she helps her mother with decisions about the family.

Frequent Words for Elinor

Present Tense Verbs 'VB': [('think', 21), ('feel', 14), ('help', 9), ('know', 9), ('censure', 9)]

Past Tense Verbs 'VBD': [('felt', 22), ('thought', 18), ('knew', 11), ('resolved', 3), ('fell', 3)]
Past Participle Verbs 'VBN': ['obliged', 14), ('hurt', 6), ('mistaken', 5), ('pleased', 5), ('felt', 5)]
ADJ [('great', 22), ('much', 21), ('good', 20), ('little', 15), ('grateful', 14)]

Words from the data produced that may be relevant to Elinor's persona are: think, help, know, censure, knew, obliged, mistaken, grateful. She is also hurt and pleased in the novel, but generally she is reserved in expressing those feelings. The most relevant words would be think and censure, because she is the sensible one in the family and attempts to guide them.

As seen in figure 2, there is a general increase in the polarity count for Elinor until around Chapter 10 when there is a decrease. This decrease continues with fluctuations until after Chapter 25 when there is a steady increase, with occasional fluctuations. The increase after Chapter 25 is most likely due to the fact that Elinor is on better terms with her sister and after being somewhat upset at the discovery of Edward's engagement, recovers.

4.3 Character 3: John Willoughby from *Sense and Sensibility*

John Willoughby, who is mostly referred to as Willoughby, is a potential love interest to Marianne. He meets the family when they move after Mrs. Dashwood's husbands death. After charming Marianne, he leaves town suddenly and when seen again he is with another women who he plans to get engaged to.

Frequent Words for Willoughby

Present Tense Verbs 'VB': [('feel', 6), ('doubt', 6), ('smile', 6), ('hope', 6), ('believe', 6)]
Past Tense Verbs 'VBD': [('thought', 10), ('felt', 8), ('obliged', 2), ('knew', 2), ('understood', 2)]
Past Participle Verbs 'VBN': [('thought', 6),('welcome', 4),('obliged', 2), ('mistaken', 2),('supposed', 2)]
ADJ: [(u'little', 7), ('good', 5), ('smile', 4), ('delightful', 4), ('full', 4)]

Words from the data produced that may be relevant to Willoughby's persona: feel, smile, hope, believe, obliged, welcome, delightful. Although these do not have to with Willoughby's actions but rather the way that others feel about him, doubt and mistaken could be relevant.

As seen in Figure 3, there is an increase in the polarity count for Willoughby. There is a small dip around chapter 30, a plateau followed by a small peak around chapter 45. The small dip occurs when Marianne finds out about his upcoming engagement. The peak occurs when he apologizes to Elinor for the way that he treated her sister and asks Elinor to relay it to the very sick Marianne.

4.4 Character 4: Edward Ferras from *Sense and Sensibility*

Edward Ferris is the older brother to Robert Ferris and Fanny Dashwood, the wife of Elinor and Marianne's half brother. He develops a close relationship with Elinor before she and her family move. Around Chapter 20, Elinor discovers that he has been secretly engaged prior to their friendship forming.

Frequent Words for Edward

Present tense Verbs 'VB': [('know', 8), ('think', 7), ('hope', 6), ('surprise', 6), ('love', 4)]

Past Tense Verbs 'VBD': [('thought', 12), ('felt', 11), ('knew', 5), ('stood', 2), ('fell', 2)]

Past Participle Verbs 'VBN': [('mistaken', 3), ('divided', 3), ('resolved', 3), ('obliged', 2), ('pleased', 2)]

ADJ [('good', 16), ('great', 14), ('sure', 11), ('possible', 8), ('much', 7)]

Words that may be relevant to Edward's persona aww know, think, thought, knew, and divided. Other words that could be relevant on some level are hope, love, mistaken, pleased. Edward is similar to Elinor in that he is a thinker, which explains why many of the words frequently associated with him have to do with knowledge or thought.

As seen in Figure 4, there is a general increase in the polarity for Edward. There is a significant dip around chapter 20 which correlates to when Elinor finds out that Edward is engaged to someone else.

4.5 Character 5: Colonel Brandon from *Sense and Sensibility*

Colonel Brandon is a friend of the extended Dashwood family who meets Mrs. Dashwood and her daughters when they first move. He is very interested in Marianne from the beginning, but she finds him awkward and boring.

Frequent Words for Brandon

Verbs that are 'VB' [('know', 3), ('wish', 3), ('think', 3), ('secure', 2), ('likewise', 2)]

Past Tense Verbs that are 'VBD' [('thought', 6), ('felt', 5), ('knew', 2), ('resolved', 1), ('stood', 1)]

Past Participle Verbs that are 'VBN' [('hurt',4), ('obliged',2), ('disinterested',2), ('established',2), ('excited',2)]

ADJ [('much', 8), ('great', 7), ('anxious', 7), ('particular', 6), ('good', 4)]

The words from the data produced that may be relevant to Brandon's persona are: wish, secure, felt, established. The words obliged and anxious may also be relevant as he is very polite and also anxious about what Willoughby might do to Marianne.

As seen in Figure 5, there is a negative polarity associated with Brandon from early on in the novel. After chapter 30, his polarity steadily increases. This increase occurs after Marianne discovers that Willoughby is not going to marry her. Elinor and Brandon become close friends and Marianne takes up an interest in Brandon, resulting in an increase in polarity.

4.6 Character 6: Emma Woodhouse from *Emma*

Emma Woodhouse, the protagonist of the novel *Emma*, is about twenty years old. She takes care of her father's household because her mother has passed away and her sister has been married off.

Frequent Words for Emma

Present tense Verbs 'VB' [('think', 35), ('help', 27), ('wish', 15), ('know', 14), ('feel', 11)]

Past Tense Verbs 'VBN' [('thought', 58), ('felt', 32), ('knew', 21), ('fell', 4), ('obliged', 3)]

Past Participle Verbs 'VBN' [('obliged', 16), ('pleased', 11), ('thought', 8), ('delighted', 6), ('felt', 5)]

ADJ[('good', 34), ('great', 34), ('little', 30), ('much', 26), ('large', 21)]

The words from the data produced that are most relevant to Emma's character are think, help, wish, feel, pleased, and delighted. Emma believes that she is helping others even if there are selfish reasons behind her actions. She also does a lot of thinking, planning, and believes herself to be very clever.

As seen in Figure 6, there is a steadily increase in polarity for Emma. Readers of the novel believe that although Emma starts off fairly selfish, she does grow in personality throughout the novel.

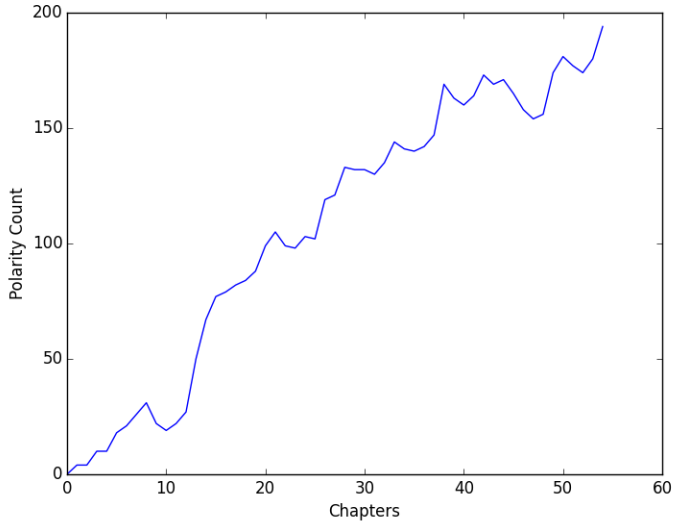

Figure 6: Graph of Polarity for Emma throughout the novel *Emma*

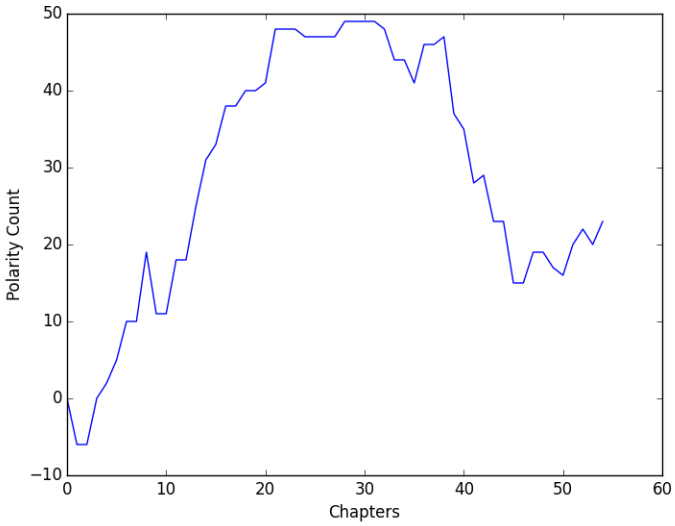

Figure 7: Graph of Polarity for Knightley throughout the novel *Emma*

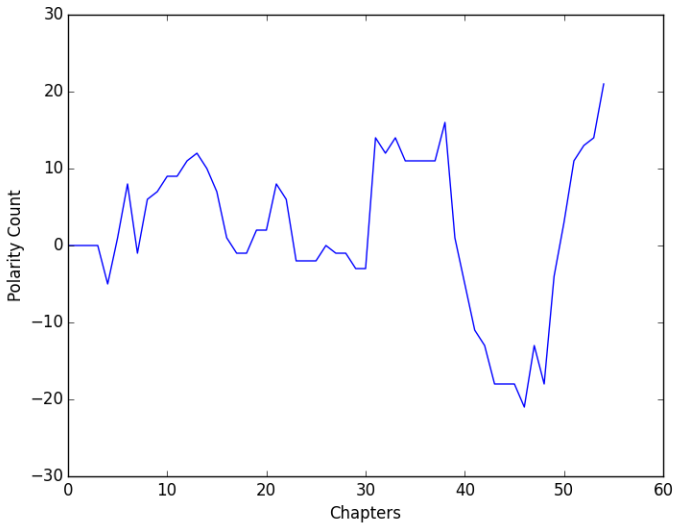

Figure 8: Graph of Polarity for Harriet throughout the novel *Emma*

4.7 Character 7: Mr. Knightley from *Emma*

Mr. George Knightley is Emma's brother-in-law. He is a friend to the Woodhouse family, but Emma doesn't always see eye to eye with him. He does not become a love interest to Emma until toward the end of the novel.

Frequent Words for Knightley

Present tense Verbs 'VB' [('think', 17), ('know', 10), ('quarrel', 9), ('mean', 6), ('consider', 5)]

Past Participle Verbs 'VBN' [('thought', 22), ('knew', 12), ('felt', 8), ('cut', 4), ('lost', 2)]

Verbs that are tagged as 'VBD' [('thought', 8), ('obliged', 3), ('delighted', 3), ('need', 3), ('pleased', 3)]

ADJ [('great', 21), ('good', 20), ('little', 19), ('much', 15), ('large', 12)]

The words most associated with Mr. Knightley from the data are: think, know, quarrel, consider and felt. Mr. Knightley is not a mean person, nor does he quarrel to be mean but often to show Emma her faults.

As seen in Figure 7, there is general increase in polarity for Mr. Knightley followed by a general decline after chapter 30. Around Chapter 45, the polarity for Mr. Knightley begins to increase again.

4.8 Character 8: Harriet from *Emma*

Frequent Words for Harriet

Verbs that are 'VB' [('think', 22), ('hope', 9), ('know', 8), ('resolve', 6), ('suspect', 6)]

Verbs that are 'VBN' [('thought', 26), ('felt', 10), ('knew', 6), ('fell', 4), ('resolved', 2)]

Verbs that are 'VBD' [('thought', 14), ('obliged', 7), ('pleased', 5), ('agitated', 4), ('glad', 4)]

ADJ [('good', 31), ('great', 26), ('poor', 19), ('much', 18), ('little', 15)]

The words most associated with Harriet that may be relevant to her persona are hope, obliged, pleased, poor and little. Harriet is a minor character compared to the other that were studied and there aren"t any words that stand out as particularly relevant. Harriet is hopeful at the prospect of marriage, she is happy to be friends with Emma, despite being poor.

As seen in Figure 8, there is a lot of fluctuation in the polarity for Harriet. There is a significant dip in polarity around Chapter 38. After Chapter 45, there is a significant increase in polarity. This dip may be related to Emma's unhappiness with Harriet and the rise may be about her and Harriet becoming closer again.

4.9 Comparative Analysis

When asked to describe Marianne, two students used the following words, "over-dramatic", "slightly", "sensitive". The students both said that "Hope" was the most relevant word from the frequently associated words for Marianne. One was surprised that the word "think" was frequently associated with Marianne. Neither were surprised by her polarity graph.

When asked to describe Elinor, the students used the following words: "thoughtful", "reserved", "prudent". They both believed that of the words frequently associated with Elinor, "think" was very relevant. One said it was interesting to have censure associated with her. They both agreed that "felt" was a surprise, and they also did not understand why her polarity graph was so negative in the middle of the novel.

Willoughby was described as "deceitful" and "selfish". Both believed that the words associated with him were relevant. However, one of the students believed that the words associated with him were more

about what other characters thought of him and not necessarily words she would use to describe him. They were both surprised that his polarity count was high at the end of the novel.

Brandon was described as “awkward”, “generous”, and “sensitive”. The data did not surprise either user and the one student believed that secure best described Brandon.

Edward was described as a “push-over”, “dishonest”, and “kind”. The data did not surprise either user and one student believed that divided was very descriptive of his person during the course of the novel, but not at the end.

Emma was described as “selfish”, “generous”, “good-hearted”. Both students felt that the description was relevant to her character but one was confused why the word “fell” was associated with her and the two other females in the other novel.

Mr. Knightley was described as “prudent”, “intelligent”, “loving”. One student thought that the descriptions matched his character and that quarreling was very much associated with his relationship with Emma. The other student was confused as to why “cut” and “lost” were associated with his character. Neither student was surprised at the polarity graph.

Harriet was described as a “push-over”, and “clingy”. One student said that her polarity graph wasn’t surprising, while the other said that it only made sense for the second half of the novel.

5 Evaluation

5.1 *Sense and Sensibility* Data

Prior to looking at the data, from reading the novel, I would describe Marianne as a dreamer, who doesn’t think about her actions before she acts. She is a very emotional character and is either very happy or very hurt. She does not read other people well and is very mistaken about their personas.

From the frequent verbs associated with Marianne, this description is supported. ‘Hope’ is the most frequent present tense verb associated with Marianne’s name throughout the novel. It is associated with her twelve times. The words thought and felt are also associated with her, which is relevant because her emotions are obvious throughout the novel. She is either happy, which is shown with the frequent association

of pleased, or sad, which is shown with the frequent association of hurt.

Marianne becomes very depressed when her lover interest leaves town. She is observed by her family to be distressed and emotional after his departure. The decrease in her polarity that was mentioned earlier is observable in Figure 1. It is obvious from the decrease in polarity that the negative words associated with her name support that emotion in the plot line.

Elinor is someone that would think before she would act. She is the cool head within her family and the voice of reason. She is not an emotional character by any means is generally reserved.

The most frequent present tense verb associated with Elinor is 'think', which is associated with her 21 times throughout the novel. This is extremely relevant to her character. Many readers associated Elinor as the brains or the 'sense' part of the novel while Marianne is associated with the 'sensibility' or emotions.

To account for the negative polarity that Elinor has in the middle of *Sense and Sensibility*, I believe that it was because of existing tensions between her and her sister. This is similar to the notable tensions that existed between Mr. Darcy and Elizabeth Bennet in *Pride and Prejudice* that was discussed in the Introduction. In other words, there are negative words associated with Elinor because of the disagreements between her and her sister. This negative polarity is not reflective of Elinor's overall persona.

Willoughby is a very charismatic character who is loved by all who meet him. Many are infatuated with Willoughby and he is what some would call the life of the party. He is somewhat deceptive, leading others on the a somewhat fake personality.

The most frequent present tense words associated with Willoughby is feel and doubt, which is interesting because 'feel' is something that his character does often. However, Willoughby himself does not doubt others. Elinor is doubtful of Willoughby's character, especially when her polarity is very negative in the novel. This association of doubt is what others feel about Willoughby. He is a character

An issue that may affect the data of these personas would be if there are two character names in one sentence. This issue was first discussed in the Example from the Introduction with Elinor and Mrs. Dashwood. If a character is mentioning another character and both names are in the sentence then all the adjectives and verbs within that sentence will be associated with both character names. Therefore, if someone is an evil character but talks very positively about other people, it could affect their overall polarity. This may have

influenced the polarity for Willoughby, who is very charismatic.

Attempting to normalize the polarity of the main characters did not result in accurate data and this could be due to the issue of words being associated with a character's name that have to do with another character entirely.

Edward is similar to Elinor in that he is very reserved and more 'sense' than 'sensibility'. He is quiet and kind. Despite not being forthcoming to his engagement when he befriends Elinor, he is overall a genuine character, and redeems himself for being awkward and distant from Elinor.

The most frequent words associated with Edward have to do with thinking, or sense. Know, think, thought, and knew are all telling of his personality. He was mistaken in his feelings for the woman that he first was engaged to and readers can see that he is very divided about what he should do because of the engagement.

The polarity graph corresponds to the plot line in relation to Edward, where there is a dip in the polarity count when Elinor finds out about his engagement from his fianc. It also shows a steady increase after Elinor discovers this which corresponds to her growing feelings for him and their marriage.

Brandon is a kind, wise man who is obviously infatuated with Marianne. He is a well-established, generous person who is very helpful and considerate. He is kind to Marianne despite knowing that she does not return his interest.

The most frequent words associated with Brandon are wish, secure, felt, established. Readers can sense that Brandon is wishful about developing a relationship with Marianne, even if he doesn't out right say it himself. Being an older man, he is secure financially and is established. He is someone who feels strongly for Marianne and felt very strongly that Willoughby was indecent. The words obliged and anxious may also refer to his feelings about what Willoughby might do to Marianne and what Willoughby had done to other girls in the past.

The polarity graph is lower than expected in the beginning of the novel. This may be due to the few times that Colonel Brandon's name appears in the beginning of the next, and also because he is a fairly neutral person. The increase in polarity occurs as expected when Elinor and Brandon become friends after Willoughby gets engaged to another. Elinor and Brandon are able to become friends and it becomes clear

that he is a loyal, kind-hearted man to the audience. His polarity also increases at the end when he marries Marianne.

5.2 *Emma* Data

In *Emma*, the main character named Emma starts out as selfish. Her mother passed away and her older sister is married so she is the only woman in the household. It is obvious that running her father's household has given Emma a lot of confidence in her abilities. Emma believes that she is helpful, she tries to make love matches to people who would likely never marry and she is completely unaware of other's affections toward her.

Words that were closest to her persona that was highlighted from the data produced for her persona were: think, help, wish, feel, pleased, and delighted. Emma believes that she is a clever person, especially when she tried to match her friend Harriet to someone who, unbeknownst to Emma, is interested in marrying her and not Harriet. She is wishful in trying to elevate Harriet's status to keep her close by, despite Harriet's wish to marry someone else.

As mentioned earlier, Figure 6 illustrates a steadily increase in polarity for Emma, which correlates the growth the Emma has in her personality, thanks in part to Mr. Knightley who points out the error of her ways and her misjudgments to her in order for her to correct her mistakes.

Mr. Knightley is an intelligent, well-intentioned man that is close with Mr. Woodhouse. He is interested in the well being of Emma Woodhouse and is often disagreeing with her about her actions. He is not afraid to point out her flaws.

The words most associated with Mr. Knightley from the data are: think, know, quarrel, consider and felt. Mr. Knightley is intelligent, so think correlates to that persona. He is often quarreling with Emma bout actions that she has done that affect others negatively. Mr. Knightley is not a mean person though. He is considerate and he does like Emma, which is especially obvious at the end of the novel when they become engaged.

As mentioned earlier, Figure 7 shows an increase in polarity for Mr. Knightley followed by a general decline after chapter 30. Around Chapter 45, the polarity for Mr. Knightley begins to increase again. I

believe that this increase occurs because Emma and Mr. Knightley are not disagreeing as much as they were, or because Emma is slowly realizing her feelings for him.

Harriet is a gullible, uneducated girl who is friends with Emma. She is very kind and sweet. She was raised without knowing her parentage and is hopeful when Emma believes that she can find a good marriage match for her.

The words most associated with Harriet from the data were: hope, obliged, pleased, poor and little. Harriet is hopeful at the prospect of marriage, she is polite and pleased at becoming part of Emma's social circle. Some of the data produced did not line up with her character but it could be due to the amount of times she is mentioned in the novel.

As seen in Figure 8, there is a lot of fluctuation in the polarity for Harriet. Again, this may be due to the face that she is not mentioned often. Another reason that there might be a lot of fluctuation is because she is a somewhat neutral person. There is a significant dip in polarity around Chapter 38, this could be because of her limited interaction with Emma, or with Emma's displeasure in her. After Chapter 45, there is a significant increase in polarity. This increase could be related to Emma and Harriet becoming closer and talking more.

6 Conclusion

The goal of this project was to develop a tool that generated a character's persona. To generate the persona, I looked at the adjectives and verbs that were associated with a character's name, produced the most frequent of these for the user, and also showed the user the rating for the character based on these words over the course of a novel.

Through examining *Sense and Sensibility* and *Emma*, novels by Jane Austen, the data produced by the program seems very promising. For each of the characters, there were words that were produced that were relevant to their persona. The polarity count for the character did not always follow their persona, however, there was a relationship between the plot line and the change in polarity.

This program will not display accurate results with novels that use more pronouns. That amount of pronouns used must be as insignificant as possible, or the number of characters in the novel must be very

few. Future work can be done to make the words more accurate for the character. This work could include deciding which words relate to a character if there are two characters in one sentence, or adding the words associated with a character if the sentence only contains pronouns. If this data is collected, it would be interesting to see if normalizing the polarity results would improve the data.

7 Bibliography

C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text based emotion prediction." *Handbook of personality: Theory and research (2nd ed.)*. New York: Guilford (1999).

D. Bamman, B. OConnor, and N. A. Smith, "Learning latent personas of film characters." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria: Association for Computational Linguistics (2013). pp. 352–361.

N. Chamber and D. Jurafsky, "Unsupervised Learning of Narrative Schemas and their Participants." *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*. Suntec, Singapore (2009). pp. 602–610

M. Elsner, "Character- based kernels for novelistic plot structure." *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.* Avignon, France: Association for Computational Linguistics (2002). pp 634–644,

F. Mairesse and M. Walker, "Automatic recognition of personality in conversation." *Proceedings of the Human Language Technology Conference of the NAACL*, Companion Volume: Short Papers. New York, New York (2006). pp.85–88

T. Miaskiewicz, T. Sumner and K. A. Koza, "A Latent Semantic Analysis Methodology for the Identification and Creation of Personas" *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Florence, Italy (2008). pp. 1501–1510

L. Mooneyham, “Romance, Language, And Education In Jane Austen’s Novels”. New York: St. Martin’s Press, 1988. Print. pp. 45-68.

O. P. John and S. Srivastava, ”The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives”, *Handbook of personality: Theory and research*. University of California (1999). pp. 102–138

J. Wiebe and E. Riloff, ”Creating subjective and objective sentence classifiers from unannotated texts.” *Computational Linguistics and Intelligent Text Processing*. (2005). pp. 486–497

T. Wilson, J. Wiebe, and P. Hoffman, ”Recognizing contextual polarity in phrase level sentiment analysis.” *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics (2005). pp. 347–354.

H. Yu and V. Hatzivassiloglou, ”Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences.” *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics (2003). pp. 129–136.

Zunshine, Lisa. Why We Read Fiction: Theory of Mind and the Novel. Columbus: Ohio State UP, 2006.

http://www.merriam-webster.com/dictionary/persona

8 Appendix

A) To explore a book from Project Gutenberg, go to my ExploreBook.py file. Within the python file, you can change the book name. For instance, for *Emma* the code is:

```
myBook = nltk.corpus.gutenberg.sents('austen-emma.txt')
```

If you were to change the words inside the quote to,

```
'austen-sense.txt'
```

then you would get the novel *Sense and Sensibility*

It is also important to change the character names in the main function of ExploreBook.py.

B) The existing Subjectivity Lexicon (Wilson et al., 2005) had more information for each word than was necessary for this project. I have extracted the word, pos tag and polarity with my code MPQAMaker.py. I have also changed the polarity from positive, negative, both, and none to 1, -1, 0 and 0 respectively. I have removed any words that did not have a polarity.