**Sample Console Output of Machine Learning Tests**

This file contains sample console output of the machine learning test modules. All console output is in bold.

# non-n-gram-only-tests.py

This is the module for carrying out purely tests not using n-grams as a feature. This is the sample output:

**The default file name is new\_formality\_data.csv**

**If this is the name of the data file, press enter**

**Otherwise, please provide the correct name, then press enter:**

**Thank you. You have confirmed that the data file name is correct: new\_formality\_data.csv**

**No of records uploaded: 7032**

**You can add the following features:**

**1 - Sentence ID**

**2 - HIT ID**

**3 - Formality**

**4 - Informativeness**

**5 - Implicature**

**6 - Length in Words**

**7 - Length in Characters**

**8 - F-score**

**9 - I-score**

**10 - Lexical Density**

**11 - Number of adjectives**

**12 - Number of verbs**

**13 - Number of adverbs**

**14 - Number of conjuctions**

**15 - Number of nouns**

**16 - Number of pronouns**

**17 - Number of modal verbs**

**18 - Number of prepositions**

**19 - Number of determiners**

**20 - Number of commas**

**21 - Number of exclamation marks**

**22 - Number of full stops**

**23 - Number of question marks**

**24 - Number of existential theres**

**25 - Number of proper nouns**

**26 - Number of capitalised words**

**27 - Number of interjections**

**28 - Average number of syllabels per word**

**29 - Average word length**

**30 - Number of NLTK stop words**

**31 - Number of words with > seven characters**

**32 - Number of words with < 5 characters**

**33 - Average word frequency**

**34 - Number of words in 35 most common words in corpus**

**35 - VADER sentiment score**

**No features have been selected yet**

**Please choose the number of a feature you wish to add and then press 'enter': 35**

**You have just selected: VADER sentiment score**

**You can add the following features:**

**1 - Sentence ID**

**2 - HIT ID**

**3 - Formality**

**4 - Informativeness**

**5 - Implicature**

**6 - Length in Words**

**7 - Length in Characters**

**8 - F-score**

**9 - I-score**

**10 - Lexical Density**

**11 - Number of adjectives**

**12 - Number of verbs**

**13 - Number of adverbs**

**14 - Number of conjuctions**

**15 - Number of nouns**

**16 - Number of pronouns**

**17 - Number of modal verbs**

**18 - Number of prepositions**

**19 - Number of determiners**

**20 - Number of commas**

**21 - Number of exclamation marks**

**22 - Number of full stops**

**23 - Number of question marks**

**24 - Number of existential theres**

**25 - Number of proper nouns**

**26 - Number of capitalised words**

**27 - Number of interjections**

**28 - Average number of syllabels per word**

**29 - Average word length**

**30 - Number of NLTK stop words**

**31 - Number of words with > seven characters**

**32 - Number of words with < 5 characters**

**33 - Average word frequency**

**34 - Number of words in 35 most common words in corpus**

**You have previously added the following features:**

**VADER sentiment score**

**Please choose an additional feature and press 'enter'**

**or press C then 'enter' to select your classifier: 31**

**You have just selected: Number of words with > seven characters**

**You can add the following features:**

**1 - Sentence ID**

**2 - HIT ID**

**3 - Formality**

**4 - Informativeness**

**5 - Implicature**

**6 - Length in Words**

**7 - Length in Characters**

**8 - F-score**

**9 - I-score**

**10 - Lexical Density**

**11 - Number of adjectives**

**12 - Number of verbs**

**13 - Number of adverbs**

**14 - Number of conjuctions**

**15 - Number of nouns**

**16 - Number of pronouns**

**17 - Number of modal verbs**

**18 - Number of prepositions**

**19 - Number of determiners**

**20 - Number of commas**

**21 - Number of exclamation marks**

**22 - Number of full stops**

**23 - Number of question marks**

**24 - Number of existential theres**

**25 - Number of proper nouns**

**26 - Number of capitalised words**

**27 - Number of interjections**

**28 - Average number of syllabels per word**

**29 - Average word length**

**30 - Number of NLTK stop words**

**31 - Number of words with < 5 characters**

**32 - Average word frequency**

**33 - Number of words in 35 most common words in corpus**

**You have previously added the following features:**

**VADER sentiment score**

**Number of words with > seven characters**

**Please choose an additional feature and press 'enter'**

**or press C then 'enter' to select your classifier: 29**

**You have just selected: Average word length**

**You can add the following features:**

**1 - Sentence ID**

**2 - HIT ID**

**3 - Formality**

**4 - Informativeness**

**5 - Implicature**

**6 - Length in Words**

**7 - Length in Characters**

**8 - F-score**

**9 - I-score**

**10 - Lexical Density**

**11 - Number of adjectives**

**12 - Number of verbs**

**13 - Number of adverbs**

**14 - Number of conjuctions**

**15 - Number of nouns**

**16 - Number of pronouns**

**17 - Number of modal verbs**

**18 - Number of prepositions**

**19 - Number of determiners**

**20 - Number of commas**

**21 - Number of exclamation marks**

**22 - Number of full stops**

**23 - Number of question marks**

**24 - Number of existential theres**

**25 - Number of proper nouns**

**26 - Number of capitalised words**

**27 - Number of interjections**

**28 - Average number of syllabels per word**

**29 - Number of NLTK stop words**

**30 - Number of words with < 5 characters**

**31 - Average word frequency**

**32 - Number of words in 35 most common words in corpus**

**You have previously added the following features:**

**VADER sentiment score**

**Number of words with > seven characters**

**Average word length**

**Please choose an additional feature and press 'enter'**

**or press C then 'enter' to select your classifier: C**

**The classifiers are:**

**1 - Support Vector Machine**

**2 - Logistic Regression**

**3 - Multinomial Bayes**

**4 - Random Forest**

**Please choose a classifier by typing a number between 1 and 4 and then press 'enter': 1**

**You have selected Support Vector Machine**

**RESULTS SUMMARY**

**---------------**

**Feature(s) tested: 'VADER sentiment score', 'Number of words with > seven characters' and 'Average word length'**

**Classifier: Support Vector Machine**

**Total predictions: 1406**

**TRUE POSITIVES: 450**

**FALSE POSITIVES: 137**

**TRUE NEGATIVES: 607**

**FALSE NEGATIVES: 212**

**Accuracy: 0.75**

**Precision: 0.77**

**Recall: 0.68**

**False positive rate: 0.18**

**AUC: 0.82**

**Balanced accuracy: 0.75**

**Process finished with exit code 0**

# n-gram-only-tests.py

This is the module for running tests purely using n-grams as a feature:

**The default data file name is new\_formality\_data.csv**

**If this is the name of the data file, press enter**

**Otherwise, please provide the correct name, then press enter:**

**Thank you. You have confirmed that the file name is correct: new\_formality\_data.csv**

**No of records uploaded: 7032**

**The n-gram types are:**

**1 - Unigram**

**2 - Bigram**

**3 - Trigram**

**4 - Unigram and bigram combined**

**5 - Unigram, bigram and trigram combined**

**Choose an option by typing a number between 1 and 5 and then press 'enter': 1**

**The representation options are:**

**1 - Binary**

**2 - Non-Binary**

**3 - TF-IDF**

**Choose an option by typing a number between 1 and 3 and then press 'enter': 1**

**The stop word options are:**

**1 - Include stop words**

**2 - No, do not include stop words**

**Choose an option by typing either 1 or 2 and then press 'enter': 2**

**The classifiers are:**

**1 - Support Vector Machine**

**2 - Logistic Regression**

**3 - Multinomial Bayes**

**4 - Random Forest**

**Please choose a classifier by typing a number between 1 and 4 and then press 'enter': 4**

**You have selected Random Forest**

**TEST SUMMARY**

**------------**

**unigram with binary representation and without stop words**

**You chose the following classifier: Random Forest**

**Please be patient - the program may take a while to run.**

**RESULTS SUMMARY**

**---------------**

**Feature tested: unigram with binary representation and without stop words**

**Classifier: Random Forest**

**Total predictions: 1406**

**TRUE POSITIVES: 436**

**FALSE POSITIVES: 87**

**TRUE NEGATIVES: 657**

**FALSE NEGATIVES: 226**

**Accuracy: 0.78**

**Precision: 0.83**

**Recall: 0.66**

**False positive rate: 0.12**

**AUC: 0.86**

**Balanced accuracy: 0.77**

**Process finished with exit code 0**

# n-gram-and-non-ngram-tests-combined.py

This is the module for running tests combining n-grams and other features (such as the number of verbs in sentences):

**The default data file name is new\_formality\_data.csv**

**If this is the name of the data file you are using, simply press enter**

**Otherwise, please provide the correct name, and then press enter:**

**Thank you. You have confirmed that the existing file name is correct: new\_formality\_data.csv**

**No of records uploaded: 7032**

**The n-gram types are:**

**1 - Unigram**

**2 - Bigram**

**3 - Trigram**

**4 - Unigram and bigram combined**

**5 - Unigram, bigram and trigram combined**

**Choose an option by typing a number between 1 and 5 and then pressing 'enter': 1**

**The representation options are:**

**1 - Binary**

**2 - Non-Binary**

**3 - TF-IDF**

**Choose an option by typing a number between 1 and 3 and then pressing 'enter': 1**

**The stop word options are:**

**1 - Include stop words**

**2 - No, do not include stop words**

**Choose an option by typing 1 or 2 and then pressing 'enter': 2**

**You can add the following features to the test:**

**1 - Sentence ID**

**2 - HIT ID**

**3 - Formality**

**4 - Informativeness**

**5 - Implicature**

**6 - Length in Words**

**7 - Length in Characters**

**8 - F-score**

**9 - I-score**

**10 - Lexical Density**

**11 - Number of adjectives**

**12 - Number of verbs**

**13 - Number of adverbs**

**14 - Number of conjuctions**

**15 - Number of nouns**

**16 - Number of pronouns**

**17 - Number of modal verbs**

**18 - Number of prepositions**

**19 - Number of determiners**

**20 - Number of commas**

**21 - Number of exclamation marks**

**22 - Number of full stops**

**23 - Number of question marks**

**24 - Number of existential theres**

**25 - Number of proper nouns**

**26 - Number of capitalised words**

**27 - Number of interjections**

**28 - Average number of syllabels per word**

**29 - Average word length**

**30 - Number of NLTK stop words**

**31 - Number of words with > seven characters**

**32 - Number of words with < 5 characters**

**33 - Average word frequency**

**34 - Number of words in 35 most common words in corpus**

**35 - VADER sentiment score**

**No features have been selected yet**

**Please choose the number of a feature to add: 34**

**You have just selected: Number of words in 35 most common words in corpus**

**You can add the following features to the test:**

**1 - Sentence ID**

**2 - HIT ID**

**3 - Formality**

**4 - Informativeness**

**5 - Implicature**

**6 - Length in Words**

**7 - Length in Characters**

**8 - F-score**

**9 - I-score**

**10 - Lexical Density**

**11 - Number of adjectives**

**12 - Number of verbs**

**13 - Number of adverbs**

**14 - Number of conjuctions**

**15 - Number of nouns**

**16 - Number of pronouns**

**17 - Number of modal verbs**

**18 - Number of prepositions**

**19 - Number of determiners**

**20 - Number of commas**

**21 - Number of exclamation marks**

**22 - Number of full stops**

**23 - Number of question marks**

**24 - Number of existential theres**

**25 - Number of proper nouns**

**26 - Number of capitalised words**

**27 - Number of interjections**

**28 - Average number of syllabels per word**

**29 - Average word length**

**30 - Number of NLTK stop words**

**31 - Number of words with > seven characters**

**32 - Number of words with < 5 characters**

**33 - Average word frequency**

**34 - VADER sentiment score**

**You have previously added the following features:**

**Number of words in 35 most common words in corpus**

**Please choose an additional feature and press 'enter'**

**or press C then 'enter' to select your classifier: C**

**The classifiers are:**

**1 - Support Vector Machine**

**2 - Logistic Regression**

**3 - Multinomial Bayes**

**4 - Random Forest**

**Please choose a classifier by typing a number between 1 and 4 and then press 'enter': 4**

**You have selected Random Forest**

**TEST SUMMARY**

**------------**

**unigram with binary representation and without stop words with the following non n-gram features:**

**'Number of words in 35 most common words in corpus'**

**Your classifier is: Random Forest**

**The test may take a while. Please be patient.**

**Feature tested:**

**unigram with binary representation and without stop words with the following non n-gram features:**

**'Number of words in 35 most common words in corpus'**

**Classifier: Random Forest**

**Total predictions: 1406**

**TRUE POSITIVES: 418**

**FALSE POSITIVES: 83**

**TRUE NEGATIVES: 661**

**FALSE NEGATIVES: 244**

**Accuracy: 0.77**

**Precision: 0.83**

**Recall: 0.63**

**False positive rate: 0.11**

**AUC: 0.85**

**Balanced accuracy: 0.76**

**Process finished with exit code 0**