Classification of News Data

Vikram Ahuja(201256040)
Manas Tewari(201256031)
Asif Hussain(201226086)

What is Text Classification

Text categorization is the task of deciding whether a piece of text belongs to any of a set of prespecified categories.

Each document d can be in multiple, exactly one, or no category at all.

The task is to learn classifiers from examples which do the category assignments automatically.

A supervised learning problem.

The use of standard, widely distributed test collections has been a considerable aid in the development of algorithms for the related task of text retrieval.

Related Works

- Text Categorization with Support Vector Machines: Learning with Many Relevant Features

In this the author use many classifiers like SVM, Naive Bayes Classifiers and k-NN and shows that SVM is a better classifier.

- INCREASING ACCURACY OF K-NEAREST NEIGHBOR CLASSIFIER FOR TEXT CLASSIFICATION

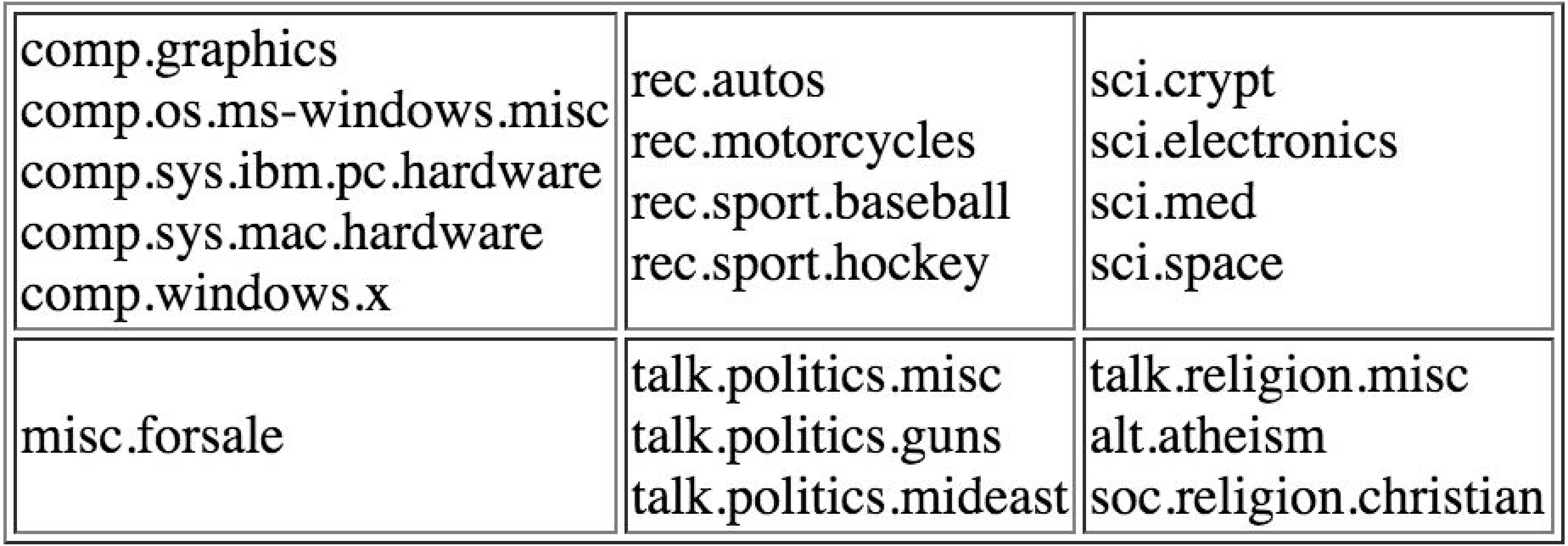
In this paper to overcome the sensitivity problem of k value by introducing a inverse cosine distance weighting voting function.

Our Approach

- Tf - idf to give scores.
- Boosting Proper Nouns
- Comparing classification algorithms like KNN, SVM and Naive Bayes both with and without boost.
- For KNN testing it with different values of distance functions like inverse cosine similarity, inverse weights(1/w), Hamming distances and levenshtein distance where possible(as dataset is huge).

Dataset

- Twenty News Group Dataset
- This dataset consists of 20000 messages taken from 20 newsgroups.
- Each newsgroup is stored in a subdirectory, with each article stored as a separate file.
- 1000 messages in each subdirectory of one topic.



comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian