SENTIMENT ANALYSIS ON TWEETS

By

Junaid Zia Khan

Ahsaan Fayyaz

Abstract

We conducted research on two topics regarding Pakistan to find out more about how the general public felt about them. We explored the issue of Kashmir through the help of python and through the use of knime.

We extensively used the twitter api, tweepy to gather tweets on the topics we wanted and then conducted the research on two platforms to get varying results, python and knime.

Python was our primary approach, and knime was the secondary approach and it played only a small part in our conclusion and overall research.

The study we did on python required much more time and effort to conduct.

We obtained scatterplots and tag clouds and accuracies through the naïve bayes classifer, all which have been detailed more in this research paper.

Motivation

Our goal was to provide sentiment analysis and really capture the opionions of the public when it came to some topics that were very controversial and were related to our country. We wanted to shed light on some topics such as terrorism, and how the people of Pakistan felt about that.

We wanted to explore our limits when it came to defining how Pakistani's felt about Jammu and Kashmir, the unmarked terrority that is still disputed to this day. Our hopes for this particular topic was to gather intel on how Paksitani's feel this area is dealt with politically. Do they want for it to be a part of the country, do they not? How do they feel about the rival country which is also clammering for a ownership of this land.

We also had other topics in mind, such as arranged marriages. And we wanted to explore what the general opinion was when it came to arranged marriages in our culture. We wanted to show the public if what we came across was negative that it is okay to venture out on your own and marry whom you want, or if it was positive we wanted to delve deeper into why it was so by conducting further research.

We also had other topics relating to Pakistan lined up such as Nawaz Sharif, our former Prime Minister, and ofcourse Bhutto.

Even though we had a plethora of topics to conduct research on, seeing as how we were pressed for time we conducted research on only a handful.

Other research papers on sentiment analysis

We have read other pieces of literature on sentiment analysis. The research conducted by Pooja Kumari and Shikha Singh (Kumari) was as follows.

They firstly collected tweets and mined them through a method that they did not disclose in the paper. Then they used categories such as either positive, negative, or junk to label each tweet and they did this manually. The junk category described a tweet that was nonsensical to the average human.

One thing that should be pointed out is that they collected tweets in all languages, so it was a continuous stream of tweets that they were collecting, then they used google translate to translate them all into English.

The method of predicting that they used although, was a naïve bayes classifier, similar to what we used.

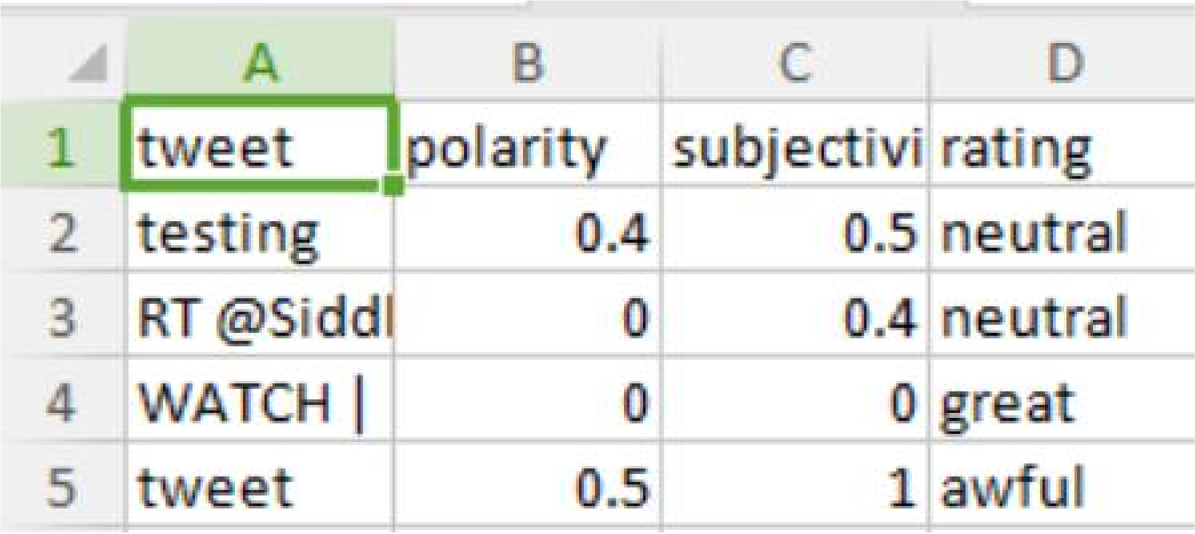
They then plotted the results in the form of a pie chart to classify if either the general consensus was that they got positive, negative, or junk tweets.

Knime approach

In our knime approach, we firstly obtained the data, through the same method that we used to populate the tables for our database that we used in the python approach. We used a python script and implemented the twitter api to gather tweets and populate an excel table for tweets on “Kashmir”.

We then wanted a way to be able to use a generative classifier, like naïve bayes in our study. So we thought, we shouldn’t use the polarity and subjectivity in our database that we got running the python script as the predicting classes, since that would be data that wouldn’t necessarily be useful to have in a naïve bayes classifier. We needed a workaround to get an accuracy.

So we set out on rating the tweets instead, on a five-point scale. We introduced five categories to measure the sentiment of the tweets, ranging from “awful”, “bad “, “neutral”, “good “, or “great”.



	A	B	C	D
1	tweet	polarity	subjectivi	rating
2	testing	0.4	0.5	neutral
3	RT @Siddl	0	0.4	neutral
4	WATCH |	0	0	great
5	tweet	0.5	1	awful

Now we had a class that we could use as our predicting class, and have the results return us something that actually made sense to a human.

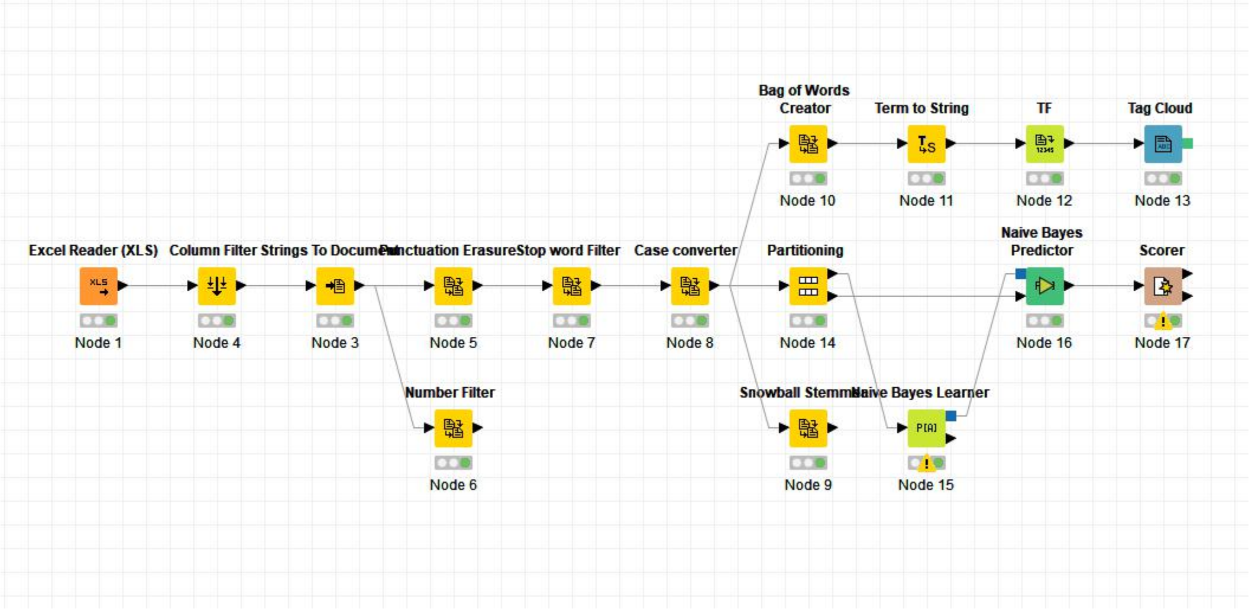
We then created the workflow for the application.

Workflow.

Firstly, we loaded the two files into xls readers using the file i/o nodes.

Then we used a column filter to filter out those two pesky “subjectivity” and “polarity” columns that were of no use to us in this part of the project.

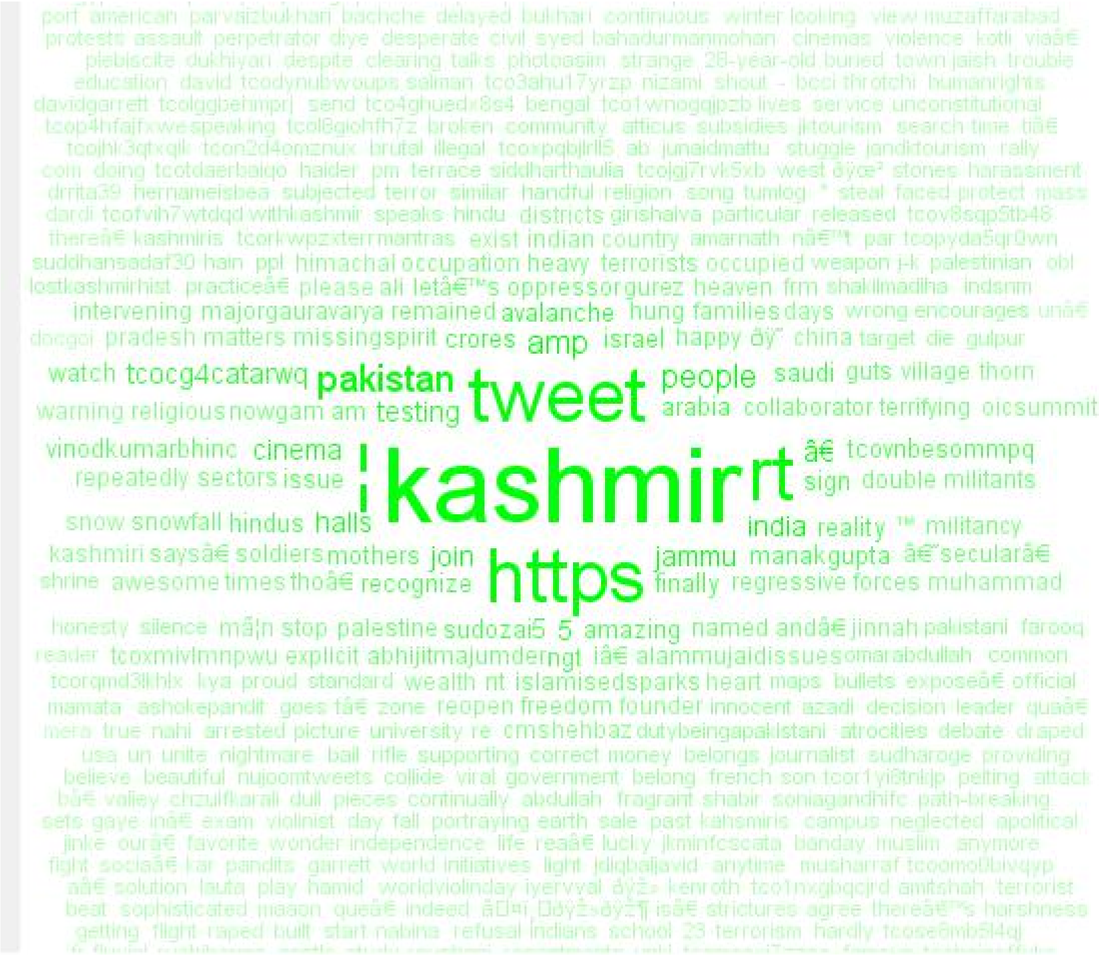
And then we created documents for each tweets, seeing as how to create a tag cloud, that is what one must do first.



We put in some preprocessing elements, such as the punctuation erasure, stop word filter, and the case converter to convert the data into an easy to read format. And then we went two routes, one to create the tag cloud and the other to run a prediction on using the naïve bayes classifier.

TAG cloud.

To create the tag cloud, first we used a bag of words creator node, then we simply created strings for each term generated, and used a Term Frequency node to measure how many times relative to each other a word was occurring. And then we had our tag cloud node which returned us this tag cloud.



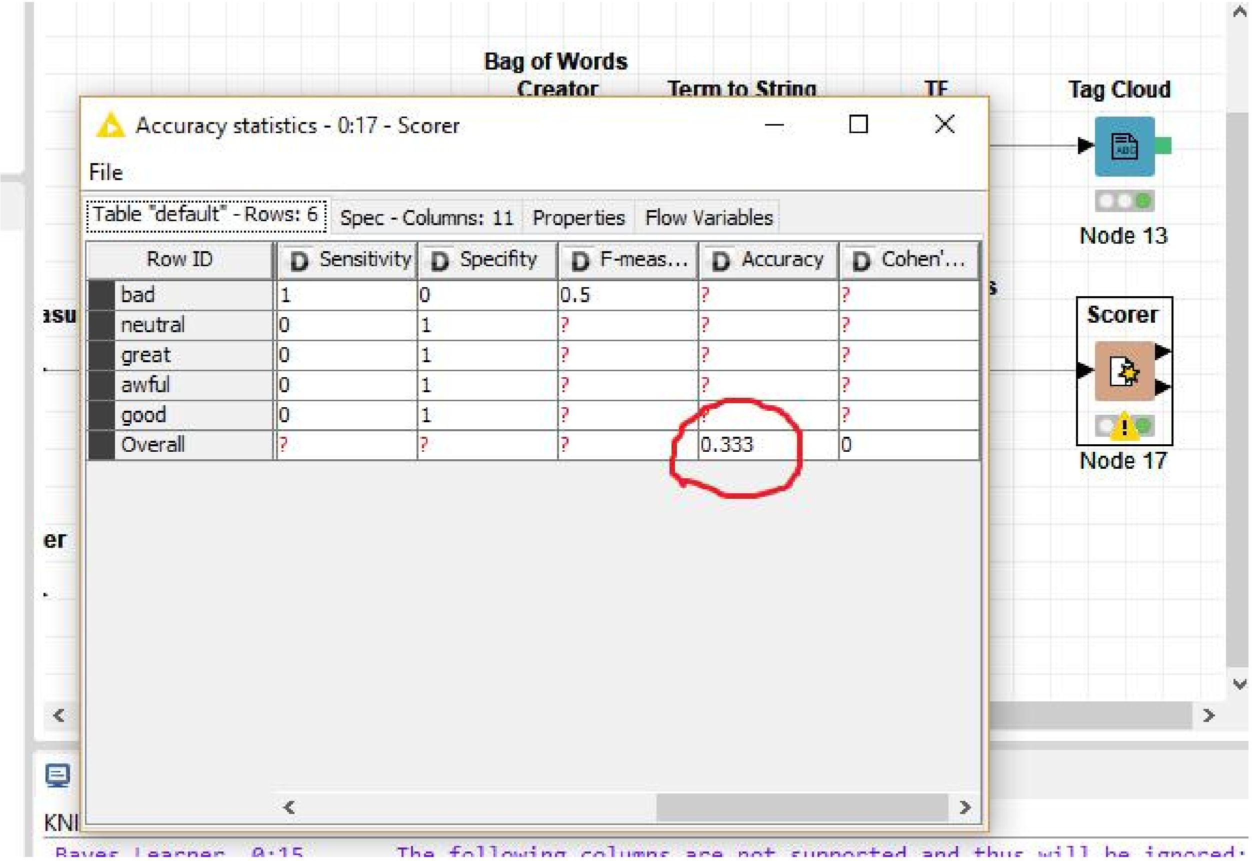
Some things that we improved since giving the presentation for this project in the knime approach were that we got the stop words node working and the punctuation erasure working, so you will see no stop words like “the” “and” “but” in the tag cloud and no punctuation.

The tag cloud most prominently returned words such as “tweet”, “pakistan”, and “issue”.

Naïve Bayes

We had partitioned the data using a 70, 30 split. That is, 70 percent of the tweets went into the learner and 30 percent went into the predictor node.

We actually got a good accuracy as far as sentiment analysis goes. The accuracy node returned us 33.3 percent as overall accuracy.



Accuracy statistics - 0:17 - Scorer

Row ID	Sensitivity	Specifity	F-meas...	Accuracy	Cohen'...
bad	1	0	0.5	?	?
neutral	0	1	?	?	?
great	0	1	?	?	?
awful	0	1	?	?	?
good	0	1	?	?	?
Overall	?	?	?	0.333	0

```
Bayes Learner   0:15          The following columns are not supported and thus will be ignored:
```

Overall, the knime approach was successful in returning a good accuracy, and generating a tag cloud that was very resourceful.

Bibliography

Kumari, P. (n.d.). Sentiment Analysis of Tweets. *IJSTE - International Journal of Science Technology & Engineering* . Retrieved from http://www.ijste.org/articles/IJSTEV1I10092.pdf